This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Social biases of mention order

Jessica Brough

A thesis submitted in fulfilment of requirements
for the degree of Doctor of Philosophy

to

Department of Psychology
School of Philosophy, Psychology, and Language Sciences
The University of Edinburgh
2022

Word count: 76,385
Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Jessica Brough
August 2022
Abstract

What can first mention, the entity mentioned first in an utterance, tell us about group identification, self-other equivalences and assumptions of agency? In the field of psycholinguistics, language is not routinely studied in relation to social identity or social contexts. Yet, language is a fundamental social behaviour that is often influenced by culture and society. Through a series of production experiments, this thesis shows that a speaker who is describing an interaction such as a hug taking place between someone from their ingroup and someone from their outgroup will demonstrate a tendency to mention the ingroup member first. Through investigating the effects of social biases in production, we also demonstrate that mention order has considerable effects on comprehension through the assumptions of agency that it creates. Thus, this thesis presents evidence establishing the importance of considering a speaker’s identity and the identities of the people they are talking about in studies of production and comprehension.

In Chapter 2, we show how relative similarity, according to gender and race, affect mention order decisions, testing two hypotheses. First, that mention order would reflect conventional social hierarchies resulting in a ‘man-first’ and a ‘white-first’ bias, and second, that speakers would show a ‘Like Me’ bias to mention the person of their own gender or of their own race first, over the person from their outgroup. In Experiments 1, 2 and 4, Black and white women and men viewed and described symmetrical interactions (like meeting or hugging) between two people who either differed in gender (one woman and one man) or in race (one Black person and one white person) and we measured who they mentioned first. Results gave some support for a man-first bias, however, they generally supported the second hypothesis. Importantly, we found a Like Me effect even when both names had the same
grammatical and semantic role, as in conjoined sentences (“Beth and Fred were hugging”), and when the second-mentioned name signified agency as in passive sentences (“Beth was being hugged by Fred”). We did not find strong evidence to suggest that these mention order biases were driven by perceptions of agency based on the visual stimuli, for instance that men were perceived as looking more like agents than women, or that people perceived the person of their own gender or race as the more likely agents (Experiment 3).

To test the effect of mention order biases on comprehension, Chapter 3 (Experiment 5) investigated how first mention influences perceptions of agency. Participants viewed transitive and conjoined sentences describing symmetrical interactions and rated who they thought was the most likely agent of the event. We found that participants tended to rate the first-mentioned name as the most likely agent. When a woman and a man’s name were paired, the order effect was modulated by gender so that the first-mentioned agent bias was stronger when it was a man’s name rather than a woman’s name.

In Chapter 4, we tested the Like Me bias further using animacy as a factor of relative similarity. We show that speakers demonstrate a robust ‘human-first’ bias when describing symmetrical interactions between a human and a robot (Experiment 6). However, when we used narrative to alter beliefs about these two groups, making the robots more similar in qualities to the speaker and the humans as less similar, non-human even, we significantly reduced this human-first effect (Experiment 7). Our findings reveal important effects of dehumanisation on language production.

The studies in this thesis demonstrate the importance of a multidisciplinary approach to understanding bias and language.
Lay Summary

Who we mention first when we are describing an interaction between two people can signify information about our social identity, the social identity of the people in the event, and the social context in which we exist. In this thesis, I have considered gender, race and human animacy as factors that are important to social cognition, investigating how they influence social biases of mention order.

The first study questioned if speakers are biased towards mentioning people of a certain gender first over people of a different gender, and towards mentioning people of a certain race first over people of a different race. I approached this question in two ways, first investigating the extent to which conventional social hierarchies, such as those that favour men over men and white people over Black people, affect who we are likely to mention first when describing a social interaction. Second, I investigated the extent to which our mention order choices are driven instead by distinctions of social similarity – whether we show a bias towards mentioning a person who is the same gender or race as ourselves first, over a person of a different gender or race. The results from this study showed some evidence for the effect of conventional social hierarchies through the discovery of a bias that saw speakers mentioning men first and women second. However, they more reliably pointed towards the second explanation – that we tend to show a ‘Like Me bias’ to mention the person who is more socially similar to us on account of being the same gender or the same race as we are.

In the second study, I considered the implications of these mention order biases, and tested if listeners or readers make judgements of agency based off who a speaker mentioned first and second in a description of an interaction. I also tested the extent to which gender affects these judgements. Results from this study showed that people show a bias towards judging the first-mentioned person as more agentive than the
second-mentioned person, and that they are more likely to do this when the first-mentioned person is a man and the second-mentioned person is a woman (e.g., Fred is hugging Beth), rather than the opposite way around (e.g., Beth is hugging Fred).

Finally, in the third study, I show that the Like Me bias can be extended to include human animacy as a property that distinguishes someone in an interaction as more socially similar to the speaker, or less socially similar. This study investigated effects of animacy on mention order by testing who a speaker would mention first when describing an interaction between a human and a robot. Results showed that speakers tended to mention the human first quite routinely, however this was greatly reduced when the robots were presented as being much more animate than the humans, through anthropomorphising the robots and dehumanising the humans.

I discuss these studies in the context of taking a multi-disciplinary approach to mention order research, that comprises of psycholinguistics, social cognition and social psychology.
Acknowledgements

The last four years have been marked by numerous experiments, plenty of analyses, and a lot of writing, but that is not what I immediately think of when I look back on this period. Instead, I think about the people I routinely sat and ate lunch with in 7 George Square throughout the working week – Anna, Ewan, Nina, Michael, Brittany, Alex and Gemma. Thank you for a friendship that will last long after my journey in higher education. I think about going around the table in lab meetings discussing our plans for the upcoming week and our biggest achievements of the previous week, in particular with Anita, Ellise, Laura, Fang, Qingyuan, Matías, Ruth, Aliff, Holly and Martin. I wish I could have met the newer lab members in person, but I have appreciated the 2D visual every Wednesday in these final years. Thank you to the whole lab group and to Julie-Anne for providing a sense of academic community, as well as reality, normalcy, and the occasional opportunity to ridicule the whole process of doing a PhD.

I do not see myself having gotten to this point without the friendships in my life. Rianna, Jade and Xine, you have shown me what dedication to your work looks like, but you have also taught me how to make speaking about your work interesting to the people who are important to you. It can be a panel discussion for your book launch, it can also be a short chat while walking along a country road, or any excuse for a writing retreat. Thank you for helping me view my research in psycholinguistics through your various disciplinary lenses, our chats about Black feminism were instrumental to me, they made this thesis better. My friends are all immensely clever – Paula, Seven, Emily, Cathy, Ben, Fionnuala, Anna, Jörundur, Kevin, Ellie, Danielle, and the whole Fringe of Colour team, it is so nice to have friends who are students, who have studied, who do not need to be at a university to keep questioning the world, and who write how I wish I
could write. Thank you to Hannah Lavery for the Writers of Colour writing group, for teaching me how to explore and enjoy writing outside of academic work.

Whenever I return to Edinburgh, there is only ever one place I want to go – to The Lighthouse on West Nicholson Street, where in the past I would leave home late on purpose, waiting for the shop to open so that I could pop in on my way to the office and start the day right. Thank you, Mairi, for all those early morning conversations, and for all the ways in which you have supported and encouraged me since I first asked you to participate in my MRes’s experiment. I am sure you had plenty of books to unbox, but you did it anyway. Who would have thought testing outside the lab would lead to a lifelong friendship? Anita, Noor, all my ex-colleagues and the whole Lighthouse family, thank you for making Edinburgh a home for me.

Holly Branigan, Lasana Harris and Hugh Rabagliati, thank you for providing me with invaluable supervision. Hugh, I will always appreciate how you agreed to supervise my MRes despite the last-minute nature of my proposal, and your willingness to take those experiments further with me through a PhD. I will also always appreciate your intuition to bring Holly into the project at the earliest stage. Holly, thank you for being a generous mentor, a collaborator, someone to run ideas past (which later turned into a PhD proposal) and a group leader. Lasana, thank you for agreeing to come onboard as my third supervisor, despite being from a totally different institution. This thesis has immensely benefitted from your insights on all things relating to social cognition, and my thinking has benefitted in general from learning from you and your lab group, and your colleagues at the EASP meeting. This thesis would also look very different without the collaboration with Emily Cross and Anna Henschel – thank you for your enthusiasm to talk to someone who really knew nothing about robots but knew some stuff about language, in order to design some cool studies about robots and
language. Thank you also for coming up with ingenious ways to keep those studies going, despite the pandemic driving us all away from the building with all the robots and to different parts of the world. A huge thank you to Alisdair Tullo for teaching me how to code so that I could build my online experiments, and to my wonderful (anonymous) models who appeared in almost every experiment.

My biggest supporters are my mum and dad, and my partner, Steisy. All the choices my parents made throughout my childhood and early adulthood meant that I was able to pursue a PhD, including allowing me to move seamlessly to Edinburgh. Thank you for always believing in me, there will be no more degrees after this. Thank you, Steisy, for encouraging and loving me so that when I moved, maybe less seamlessly, to Barcelona, I could end up in this moment, getting ready to submit this thesis, in a much more beautiful stage of my life than I could have imagined when I first started. The friends I have made in Barcelona don’t really know what it is that I do, and that doesn’t bother me at all, because they have also been very important to me in these last few years. Thank you, Luke, Tarik, Marion, Simon, Xue, Dan, Omar, Maya, Will, Jon and Marcia for adding to my life outside of the PhD, I hope we continue barely talking about work.

And, finally, thank you to Beyoncé, for dropping “BREAK MY SOUL” at the exact moment I needed it the most – when I was finalising the first full draft of this thesis.
# Table of contents

1. Literature Review .................................................................................................................. 13
   1.1. Mechanisms of Language Production ............................................................................. 15
   1.2. Mention Order and Conceptual Hierarchies ................................................................. 18
   1.3. Mind perception .............................................................................................................. 24
   1.4. Socially Biased Language ............................................................................................... 27
   1.5. Me First and Like Me ..................................................................................................... 34
   1.6. Mention Order and Comprehension ............................................................................... 38
   1.7. Thesis studies ................................................................................................................. 44
   1.8. Summary ......................................................................................................................... 47

2. Study 1, Experiment 1 – 4: Effects of gender and race on mention order and perspective .................................................................................................................. 49
   2.1. Introduction ..................................................................................................................... 49
   2.2. Background .................................................................................................................... 51
   2.3. Experiment 1 .................................................................................................................. 63
       2.3.1. Method .................................................................................................................... 63
       2.3.2. Analysis .................................................................................................................. 67
       2.3.3. Results ................................................................................................................... 74
       2.3.4. Discussion ............................................................................................................. 92
   2.4. Experiment 2 .................................................................................................................. 96
       2.4.1. Method .................................................................................................................... 98
       2.4.2. Results .................................................................................................................. 101
       2.4.3. Discussion ............................................................................................................. 117
   2.5. Experiment 3 .................................................................................................................. 119
       2.5.1. Method .................................................................................................................... 120
       2.5.2. Results .................................................................................................................. 127
       2.5.3. Discussion ............................................................................................................. 135
   2.6. Experiment 4 .................................................................................................................. 140
       2.6.1. Method .................................................................................................................... 141
       2.6.2. Results .................................................................................................................. 143
       2.6.3. Discussion ............................................................................................................. 159
   2.7. General Discussion ....................................................................................................... 160
       2.7.1. Support for main hypotheses ................................................................................... 160
2.7.2. Effects of syntactic form ................................................................. 162
2.7.3. Effects of dyad demographic ...................................................... 163
2.7.4. Intersectional effects and individual differences ....................... 164
2.7.5. Effects of agency .......................................................................... 167
2.7.6. Control and exploratory analyses ............................................. 168
2.7.7. Implications for comprehension ................................................ 169

3. Study 2: Effects of first mention on agency judgements ................. 171

3.1. Introduction .................................................................................. 171
3.2. Method .......................................................................................... 179
  3.2.1. Participants ............................................................................ 179
  3.2.2. Materials ............................................................................... 179
  3.2.3. Procedure ............................................................................ 181
3.3. Analysis ....................................................................................... 183
  3.3.1. Exclusions .......................................................................... 185
3.4. Results .......................................................................................... 186
  3.4.1. Primary analysis ................................................................. 186
  3.4.2. Gender analysis ................................................................. 190
3.5. Discussion .................................................................................... 196

4. Study 3: Animacy and human-robot interactions ............................ 203

4.1. Introduction .................................................................................. 203
4.2. Experiment 6 .................................................................................. 217
  4.2.1. Method ............................................................................. 217
  4.2.2. Results ............................................................................ 229
  4.2.3. Discussion ........................................................................ 239
4.3. Experiment 7 .................................................................................. 241
  4.3.1. Method ............................................................................. 243
  4.3.2. Results ............................................................................ 250
  4.3.3. Discussion ........................................................................ 272
4.4. General Discussion ...................................................................... 273
  4.4.1. Anthropomorphism and the animacy bias ............................. 274
  4.4.2. Attitudes towards robots and perceptions of animacy .......... 278
  4.4.3. Effects of valence, age, and a Like Me bias ......................... 282
  4.4.4. Effects of gender ............................................................... 283
  4.4.5. Memory for names ............................................................ 286
5. General Discussion

5.1. Summary of empirical findings

5.1.1. Effects of social cognition on mention order

5.1.2. Effects of valence

5.1.3. Mention order and syntactic form

5.1.4. Visual cues of agency

5.1.5. Effects of animacy on mention order

5.1.6. Effects of mention order on perceptions of agency

5.2. Explanations and implications of this research

5.2.1. Explanations for The Like Me bias

5.2.2. Explanations for the first-mentioned agency bias

5.2.3. Future directions

5.2.4. Societal implications

5.2.5. Limitations

5.2.6. Methodological contributions

5.3. Conclusion

6. References

7. Appendix A: Experimental stimuli used in Experiments 1-7

7.1. Experimental stimuli used in Experiment 1

7.2. Experimental stimuli used in Experiment 2, 3 and 5

7.3. Experimental stimuli used in Experiment 4

7.4. Experimental stimuli used in Experiments 6 and 7

8. Appendix B: Scales used for questionnaires in Experiments 1 – 7

8.1. Items Measuring In-Group Identification for gender and race

8.2. Items measuring attitudes towards same-gender relationships and interracial relationships

8.2. Items Measuring Attitudes towards Robots

8.3. Items for Manipulation Check (Dimensions of Mind Perception and Dehumanisation scale)
1. Literature Review

What a speaker mentions first in an utterance (for example, “the student colluded with the professor” versus “the professor colluded with the student”) can reveal subtle conceptual biases that influence how a listener interprets an event. Importantly, speakers do not randomly choose what they will mention first. Instead, this depends on the event structure (which entity is doing what), and it also depends on a variety of less obvious factors, such as prototypicality (Cooper & Ross, 1975; Iliev & Smirnova, 2016), imageability (Bock & Warren, 1985), relevance (Kesebir, 2017) and animacy (Branigan et al., 2008; McDonald et al., 1993; Prat-Sala & Branigan, 2000). In this thesis, I provide evidence for relative social similarity between the speaker and the people involved in the interaction as an important new factor affecting how we structure descriptions of social interactions. Across 7 experiments, I document this effect, which I refer to as “the Like Me bias”, and highlight its consequences, while investigating the factors that generate social biases in language use.

Mention order is a key component of descriptions of interactions, as the speaker must make decisions about sentence subjecthood, for instance if there is a subject and direct object such as in transitive sentences (e.g., “the student talked to the professor”), or if both referents will share the subject position, as in conjoined sentences (e.g., “the student and the professor talked”). Mention order is also an important social issue; one that can be easily overlooked but has the potential to reveal important, hidden biases of social order. This chapter will review research on mention order from different disciplines which have addressed this topic from a range of perspectives. This includes a psycholinguistic perspective, for the purpose of studying mechanisms of language production and language development, and a social cognition perspective, for the
purpose of studying how group distinctions and social biases manifest in language. In this thesis, I attempt to unify these disciplines to answer questions about language that have not yet been addressed.

First, in Section 1.1, I focus on how mention order has been studied in the context of understanding mechanisms of language production. Here, I review Levelt’s model of production and important research on incrementality, illustrating how this process underpins ordering choices. An important contribution of psycholinguistic research is the finding that production is highly influenced by conceptual factors, so much so that these are often reliable predictors of what a speaker will mention first when describing scenes or events. I discuss these conceptual factors in Section 1.2, focusing on conceptual accessibility and the animacy bias. While highly influential, this research has not yet fully considered conceptual factors that are social in nature, for instance how social identity, stereotypes and social conventions might influence mention order for descriptions of interactions between people. However, research from outside of psycholinguistics can inform our understanding of important social concepts that may also influence who a speaker will mention first. Thus, in Section 1.3, I review work from the domains of social cognition and social robotics, discussing the relevance of this work to conceptual accessibility and predictions of mention order.

Outwith psycholinguistics, other domains have more extensively considered the role of social cognition on language production. In Section 1.4, I review important research from social psychology on socially biased language. Then, in Section 1.5, I discuss how a relatively limited but interesting number of studies have addressed social biases of mention order, including identifying how these manifest in language production and considering effects on comprehension. In Section 1.6, I address comprehension more specifically, reviewing research that has shown that people make
semantic inferences from the ways in which a sentence is structured, including from the order in which entities are mentioned.

Finally, I present a roadmap of the studies in this thesis, demonstrating how each experiment aims to build on our understanding of mention order biases, including how they manifest in language production and the consequences they have for comprehension.

1.1. Mechanisms of Language Production

This thesis focuses on the third person perspective, sometimes referred to as “people watching” and relating to judgements a person makes when observing an interaction that does not involve them (Quadflieg & Penton-Voak, 2017). The studies in this thesis address how this perspective is expressed in language production and comprehension. I argue that observers are an interesting case for language researchers, as they bring with them their own impressions, stereotypes, and expectations, all of which affect their language production in some way. Furthermore, the types of interaction a person observes can vary immensely, from the valence (whether the event is a positive or a negative one, for example) or the codability of the event (whether it is easy to decipher or could be interpreted in many different ways), to the typicality of the event (if it is something regularly seen on a daily basis versus an unusual interaction) or the relevance of the event to the observer (for example, if it is something that could impact them or not). Participants involved in an interaction will also vary greatly between cases, for instance people of different ages, different social roles, or different social groups, and so descriptions of these interactions are plausibly affected by the myriad of ways in which two people engaging in an interaction differ from one another. Curiously, however, social cognition has not been a noteworthy focus across the field of
psycholinguistics. In this thesis, I aim to fill this gap and demonstrate why it is important to consider the characteristics of the speaker and of the speaker’s social context when studying mechanisms of language.

In the domain of language production, observer descriptions have been widely studied using the picture-description paradigm, in which a participant views a scene, often in a still image, and talks about it as directed. Language researchers have used this paradigm in their attempts to identify the intricacies of event descriptions, from the moment a participant first views an image to the onset and completion of their utterance (e.g., Bock & Warren, 1985; Gleitman et al., 2007; Griffin & Bock, 2000; McDonald et al., 1993; Prat-Sala & Branigan, 2000). This has provided us with some crucial information about utterance planning and production, with mention order used to pinpoint the specific mechanisms involved.

Notably, language production is often talked about as a three-stage process, beginning with conceptualisation, during which an event is extracted and conceptualised, then formulation, involving the retrieval of the linguistic information required for an utterance, and lastly articulation (Levelt, 1989; Bock & Levelt, 1994; Bock et al., 2004; Pechmann, 1989). During conceptualisation, a speaker rapidly extracts the gist of an event, recognising what is taking place and the roles within the event (Hafri et al., 2013, 2018). Then, during formulation, a speaker commences the process of grammatical encoding, choosing lexical items for concepts and a suitable syntactic structure (Konopka & Meyer, 2014), and phonological encoding, preparing to sound out these forms for the final stage of articulation (see also Garrett, 1975).

In the formulation stage, speakers retrieve the words they will need on a piece-by-piece basis as they become available, rather than waiting till they have fully prepared an utterance before they start speaking (Levelt, 1989). In this way, lexical retrieval is
considered incremental (Ferreira, 1996; Kempen & Hoenkamp, 1987; Zhao & Yang, 2016), meaning a speaker is able to begin articulating an utterance as soon as information becomes available to them (Branigan et al., 2008). Experimental evidence for this comes mostly from studies using eye-tracking with picture description tasks. Importantly, Bock and colleagues have used this paradigm to map a speaker’s conceptualisation of an event (measuring when and where they gazed while viewing the picture) and their subsequent utterance. In doing so, they determined that the early processes of production, conceptualisation and formulation, can occur in parallel (Griffin & Bock, 2000). This finding gave rise to the proposal of what is known as structural incrementality, which proposes that a speaker’s early decision-making considers the relation between entities, and that lexical information will be encoded and retrieved incrementally as their message is built up to satisfy the way in which the speaker has conceptualised that event (Bock et al., 2004; Lee et al., 2013).

However, structural incrementality is not the only strategy that speakers use during lexical retrieval. Also using eye-tracking, Gleitman and colleagues (2007) identified that speakers, who had been cued to look towards one of two entities in a scene, would show a tendency to begin their utterance with that entity, sometimes before ever gazing towards the other entity or the rest of the scene. As a result, they argued that speakers have another strategy that allows them to produce fluent utterances. What they looked at first is what they mention first, and they will continue to retrieve lemmas and syntactic information that follow from this specific starting point. This strategy, known as lexical incrementality (Brown-Schmidt & Konopka, 2015; Konopka & Meyer, 2014), ultimately allows for rapid, perception-driven utterances.

Which strategy is used and when can depend on a range of factors, including the features of the image and the complexity of the task. For example, when objects in the
event differ in perceptual salience, whereby a speaker’s attention is drawn more towards one object than another, the strategy of lexical incrementality may be employed (Gleitman et al., 2007). When they do not differ in salience, and so perceptual cues are not a determinant of where a speaker will fixate, a structurally incremental strategy may be more likely (Griffin & Bock, 2000). The ease with which the speaker can apprehend the event may also be a determining factor of which strategy is used, with speakers falling back on perceptual cues to guide their utterances when the event is high in ambiguity, but favouring structural incrementality when apprehending an unambiguous event (Kuchinsky & Bock, 2010).

1.2. Mention Order and Conceptual Hierarchies

Incrementality explains how a speaker comes to mention words in a sentence. Other important theories have revealed what a speaker will mention and when. Bock has argued that the retrieval of a concept or lemma is importantly influenced by what is known as conceptual accessibility (Bock, 1986a, 1986b; Bock & Warren, 1985; McDonald et al., 1993). Conceptual accessibility defines the conceptual relations that an entity participates in, as the number of relations determines how easily a concept can be retrieved from memory. Central to this is an entity’s predicability, defined as how many things it can do. For instance, a person can do more things than a plant, including running, sitting down, developing a theory, questioning the meaning of life, compared to the more limited abilities of a plant. It is this number of associations that can be made with the entity, including the verbs or adjectives that can be attributed to it, that increases the pathways to its retrieval and determines its relative conceptual accessibility (Tanaka et al., 2011). An entity that has more predicates associated with it will be more conceptually accessible than an entity that has fewer predicates associated
with it. And, since language is incremental, with speakers encoding lexical information as it becomes available to them, concepts that are easier to retrieve will occur earlier on in utterance (Branigan et al., 2008).

How is the conceptual accessibility of an entity determined? Prat-Sala and Branigan (2000) suggested that accessibility takes two forms: inherent accessibility and derived accessibility. Inherent accessibility describes certain fundamental characteristics of an entity, such as its animacy, prototypicality and concreteness, which are constant across different contexts. In contrast, derived accessibility refers to the temporary accessibility of an entity which may be increased by the communicative context, for example by semantic priming or discourse salience (Bock, 1986a). Taken together, an entity’s overall conceptual accessibility will be determined by both factors.

There are multiple determinants of conceptual accessibility, which have been tested experimentally. For instance, entities that are more imageable are considered more accessible than entities that are more abstract. Evidence for this comes from Bock and Warren (1985), who asked participants to recall a set of sentences in which the subject, direct object and indirect object varied in how easily their image could be conjured. During recall, participants were likely to rearrange the structure of the original sentences in a way that showed a hierarchy of conceptual accessibility. When the subject of the original sentence was less imageable than the direct object, participants tended to mention the more imageable noun first as the new sentence subject during recall. Thus, the more conceptually accessible entity appeared in a higher grammatical function than the less conceptually accessible entity. Importantly, they did not find the same re-ordering effect when participants were recalling conjunctive noun phrases (e.g., *time and winter* in “the lost hiker fought time and winter”), and consequently proposed
that it is the assignment of grammatical roles that is influenced by conceptual accessibility, rather than the sequence of first and second mention.

Furthermore, entities that are more typical to a semantic category (e.g., *apple* in the category of fruit), otherwise known as prototypicality, tend to occur earlier on in an utterance than entities that are less typical (e.g., *kiwi*, Kelly et al., 1986; Onishi et al., 2008). Prototypical entities are considered easier to learn and easier to categorise, and so they represent another means by which an object can be easier to retrieve from memory. Interestingly, during their exploration of typicality and syntax, Onishi and colleagues found that this property only reliably predicted which noun would be mentioned first in conjoined noun phrases (e.g., “she took the apple and the kiwi to school”), and not when the two nouns were ascribed different grammatical functions in a sentence (e.g., “the apple was in the bag with the kiwi”).

Widely studied for its effects on mention order and syntax choices, animacy is a particularly interesting determinant of conceptual accessibility. The more animate something is, the more predicates can be associated with it, or the more it can do, and so the easier it is to retrieve from memory (Prat-Sala & Branigan, 2000). The hierarchy of conceptual accessibility, when specifically referring to effects of animacy, has become known as the animacy hierarchy (Aissen, 2003; Harris, 1978). Entities that are more animate appear higher up than entities that are less animate, with humans at the top of the hierarchy, to non-human animals and human-like objects such as dolls, to inanimate living entities such as plants, then inanimate non-living entities such as rocks and other concrete masses, and finally abstract concepts at the bottom (Figure 1).
Figure 1

The animacy hierarchy

Note. Inanimate human-like entities refer to anthropomorphised objects such as robots and dolls with the appearance of humanness. This category is not mentioned in most discussions of an animacy hierarchy (e.g., Aissen, 2003; Harris, 1978) but is alluded to in others (see Yamamoto, 1999 for a discussion of the animacy of machines, computers and stuffed toys). I propose that entities within this category would typically hold a similar position to non-human animate entities; near the top of the hierarchy but beneath human beings.

Numerous studies spanning many decades have identified the link between animacy and subjecthood. Entities in the subject position tend to be judged as more animate than those in the object position (Johnson, 1967). Clark and Begum (1971) found sentences rated for acceptability followed a hierarchy of animacy, whereby
sentences with human subjects were most acceptable, sentences with non-human animate subjects were the next most acceptable, and sentences with inanimate subjects were least acceptable. This follows from the fact that the subjects of sentences are more often animate than inanimate (MacWhinney, 1977). The tendency for speakers to produce utterances in which the most animate entity is prioritised temporally, i.e., mentioned earlier on than less animate entities, is also known as the animacy bias (Lempert, 1989; McDonald et al., 1993, see also Branigan et al., 2008 for a review). This bias has been identified for both adult (Tanaka et al., 2011) and child speakers (Prat-Sala et al., 2000; Franz et al., 2021; Buckle et al., 2017; Buckle, 2020), for neurodiverse speakers (Lake et al., 2010) and across different languages (e.g., Branigan & Feleki, 1999; Christianson & Ferreira, 2005). Animacy effects also tend to lead a speaker towards producing a normally disfavoured syntactic structure such as a passive, in order to secure the more animate entity’s position as sentence subject (Harris, 1978; Prat-Sala & Branigan, 2000), for example, “the man was chased by the dog” over “the dog chased the man”.

Importantly, research has also shown that animacy affects mention order independently of grammatical function (Branigan & Feleki, 1999; Christianson & Ferreira, 2005). Animate entities are often mentioned before inanimate entities in frozen conjuncts (e.g., “people and things”, “man and machine”) as initially detailed by Cooper and Ross (1975). Recently, Franz and colleagues (2021) demonstrated the effect of animacy on conjuncts (e.g., “dolphin and planet”) experimentally with German-speaking preschool-aged children and adults. Their results also provided evidence that the animacy bias increases with age. Consequently, this suggests that conceptual accessibility can influence word order without the distinction of grammatical function.
While hugely influential for how we understand the effects of conceptual factors on language, to some extent the animacy bias, as it is currently understood, has placed limits on how we understand mention order biases. Studies on relative animacy have paired humans with non-humans as a condition of different-animacy and humans with other humans as a condition of same-animacy (e.g., Christianson & Ferreira, 2005). Since its inception, the animacy hierarchy has generally stopped with the highest category as *human* (apart from maybe Cooper and Ross, 1975, who considered the category of deity or the divine as above that of human). This has been presumably based on an assumption that all humans are represented in roughly the same way (although see also Yamamoto, 1999 and The Hierarchy of Persons, ordering how speakers use the first, second and third person to address people). However, research from social psychology contradicts this; social categorisation underpins and influences how we think about people (e.g., O’Doherty & Lecouteur, 2007), while perceptions of what a person can do can vary depending on stereotypes associated with their identity (Haslam, 2006). Thus, differences in the representations we hold for one person over another will be determined by dimensions of social cognition. These dimensions can, and will, affect an important social process – perspective-taking (Davis et al., 1996; Galinsky & Moskowitz, 2000), which is often signified by the order of mention.

The animacy bias allows us to make good predictions of what a speaker will mention first when describing an interaction between a person and an animal or object, but what if that interaction was between two people, whose identities create differences in how they are represented socially? To answer this question, we would have to think about everyone involved in the description: the two people engaging in the interaction, and the speaker who is describing it. In psycholinguistics, studies of language have not often considered social cognition as it relates to the social characteristics of the speaker.
and their social context. It would be difficult to say exactly why this has been the case — perhaps it is the consequence of schools of thought that have more often approached language as a formal system with models of data driving acquisition, production and comprehension, and with a focus on aspects like universal rules of linguistics and the innateness of language (Pinker et al., 1993). In this thesis, however, I take a novel approach by maintaining that considerations of social cognition are required for furthering our understanding of language mechanisms.

1.3. Mind perception

In psycholinguistics, conceptual accessibility is a fundamental theory for understanding how entities differ from each other. However, outside this field, these differences have been framed in slightly different ways. In cognitive psychology, a fundamental tenet is that there are variations in the mental abilities between entities. The ‘dimensions of mind perception’ proposed by Gray and colleagues (2007) identify two features, agency and experience, as key in a person’s perception of another entity’s mental abilities, or ability to mentalise. Here, agency is defined as the ability to plan and act, while experience is defined as the ability to sense and feel. Thus, the ability to perceive something as having a mind comprises of the ability to consider how that entity thinks, feels and acts (and why).

Gray et al. (2007) collected ratings for agency and experience for a range of entities, thus mapping their relative dimensions of mind perception. Where humans were rated as having high agency and experience, animals were rated as having low agency but high experience, and robots, which are inanimate but anthropomorphised (non-human but having human-like qualities, Bartneck et al., 2009), received relatively low ratings of agency and barely any experience. Compared to the conceptual
accessibility hierarchy and animacy hierarchy, which assume a relatively linear placement of entities, these dimensions perhaps more clearly show how entities can be conceptualised as similar in one respect but different in another.

Distinctions between animate and non-animate entities are intrinsic to social cognition. Neurological research has focused on how these distinctions manifest in the brain (Adolphs, 2010; Grossmann & Johnson, 2007), including revealing how they allow us to perceive important social cues such as agency and intention (Blakemore & Decety, 2001; Castelli et al., 2000; Heider & Simmel, 1944), emotions (Kegel et al., 2020), movement (Kaduk et al., 2013) and faces or face-like patterns (Gobbini et al., 2007). Importantly, the dimensions of mind perception demonstrate not only how entities of different animacy are perceived, but also how different types of people are perceived. Here, the categories “you”, “woman” and “man” hold similar positions on this scale, while the category of “girl” received a lower rating of agency, “baby” received a high rating of experience but very little agency, the category of a man in a “persistent vegetative state” received a low rating of agency and medium rating of experience, and the category of “dead woman” received low ratings for both mind properties.

While cognitive theories of mind perception have focused on consistencies in how we perceive minds, more recent social theories have focused on how context causes variation in mind perception. Indeed, the different categorisations of person mentioned above bring to mind another important subject matter in social psychology; that at times during person perception, expectations relating to mental abilities are flouted. Dehumanisation is commonly defined as the refusal to fully mentalise a person or group, for instance by simplifying their capacity to experience emotions or feelings, or by homogenising a group according to stereotypes rather than appreciating the
complexity of their identities (Haslam, 2006). Haslam identifies dehumanisation as comprising of two types; animalistic dehumanisation, whereby people are perceived as unrefined and more animal-like than human-like, and mechanistic dehumanisation, whereby people are perceived as cold, automated and machine-like with little agency. Animals and animated machines like robots all fall lower on the conceptual accessibility hierarchy than humans, and in different ways (Branigan et al., 2008). This is, interestingly, also represented in the pathways people take to dehumanise others. It might then be reasonable to propose that dehumanisation affects a person's predicability, if they are not associated with the same number of predicates (for instance, behaviours, emotions and abilities) as someone who is considered more fully as human.

That Gray and colleagues (2007) included robots in their study was prophetic of how robots are now being considered as potential social agents. In social robotics, markers of human animacy are often used to make robots appear more like a human (Broadbent, 2017). Research from social robotics has found that stimulus cues, those which imply humanness through visual information such as faces and movement (Cross et al., 2016), engage the social cognition brain network similarly when viewing human-like features as when viewing actual human features. For example, viewing robotic faces and human faces (Gobbini et al., 2011) or when viewing facial expressions of computerised avatars and humans (Kegel et al., 2020).

Additionally, knowledge cues, which use beliefs about an entity to provide information about its animacy, can also vary the perception of humanness of a non-human entity in a way that engages social cognition. Evidence for this comes from the finding that someone is more likely to automatically imitate an ambiguous hand wearing a glove when they had been led to believe the hand was human, rather than
wooden (Liepelt & Brass, 2010; Cracco et al., 2021). Further evidence comes from Cross and colleagues (2016), who have found that the so-called “social brain network” is more engaged when someone has received prior information that the movements of an avatar were created using human motion capture, rather than when they have been told that the movements were fully computer-generated.

In the language of conceptual accessibility, stimulus and knowledge cues would represent a robot’s inherent accessibility. Thus, when these cues suggest that the robot is capable of doing more things, they are reflecting a greater predicability of the robot than one that has fewer human-like features. This also calls back to Gray et al.’s (2007) dimensions of mind perception, because a robot that is perceived as able to plan and act, and think and feel, will have more predicates associated with it in the form of more verbs that can be attributed to it than a robot that is perceived as unable to do these things (for example, being able to say that the robot can communicate or that it can be confused). Studies on mind perception and dehumanisation remind us that not all people are conceptualised the same, while research from social robotics demonstrates that social cues are likely to affect a social being’s inherent accessibility. Thus, findings from the domains of person perception and social robotics could help answer the question of who a speaker would mention first when describing an interaction between two people. These factors are all likely to affect who, or what, a speaker chooses to talk about first.

1.4. Socially Biased Language

For the most part, psycholinguistic studies that have focused on mention order have done so for the purpose of learning about the mechanisms of language production. Research from social psychology that has addressed mention order has worked with a
different objective, namely with the goal of learning about social behaviour such as stereotyping and group associations.

Conventional social hierarchies exist across societies and differ from culture to culture, established by age, economic position, race or colourism, educational background, relationship status, gender, physicality and more. Importantly, positionality within a hierarchy (i.e., group identification) does not necessarily determine if someone will conform to its expectations, for example, people from all genders promoting misogynistic beliefs or people of different races endorsing racist rhetoric (Taylor, 2017). This makes studying social hierarchies a complex but pertinent issue.

Where do social hierarchies appear in production? In a gendered context, there are numerous examples where male nouns come before female nouns in frozen conjuncts (those which might sound unnatural if reversed, such as “king and queen”, “boys and girls”, “sons and daughters”; Cooper & Ross, 1975; Mollin, 2012). While studying gender stereotypes in everyday language from a social psychology perspective, Hegarty and colleagues (2011) tested the effect of gender on mention order. In Study 1, they looked at the number of hits from a Google search for pairs of the most common names in the UK and the USA and found that name pairs with a man’s name first and woman’s name second (e.g., “Jack and Emily”) were more common than name pairs in the other direction (e.g., “Emily and Jack”).

Effects of gender on mention order are not unidirectional, however. Kesebir (2017) argues that contextual relevance may modulate gendered word orders, such that feminine lexical items will come before their masculine counterparts when stereotypes imply that femininity is more relevant. To test this, she used a media outlet database to measure the total number of hits for a selection of conjoined noun pairs belonging either to the category of “nonkinship” (e.g., woman/man, girls/boys,
businesswoman/businessman and their reverse versions) or “kinship” (family-related nouns, e.g., mother/father, grandmother/grandfather, aunt/uncle and their reverse versions). For nonkinship words, the male lexical items came first on the majority of hits, but for kinship words, there was no statistical difference between first-mentioned female or male lexical items. Kesebir argues that this difference between nonkinship and kinship words indicates that perceived relevance of women in family roles has an influence on mention order. Where family roles are not relevant, the man will be mentioned first, but when they are, the likelihood that the woman will come first increases. This finding shows how the speaker’s social context, which can include hierarchies, stereotypes and other factors affecting how we understand the world, can influence their mention order choices. In this case, the social context is the existence of stereotypical expectations relating to gender and gender roles (Eagly et al., 2000). Thus, characteristics of the speaker’s social environment can affect who they will mention first.

Hegarty and colleagues (2011) also studied the effect of gender stereotypes on mention order in their second study. This time they took an experimental approach, asking female and male participants to imagine and name heterosexual couples who either conformed or did not conform to gender stereotypes (e.g., by doing more housework, or earning more money). Results showed that participants tended to mention the man first when naming a stereotypical couple, in which both people were said to take on their respective traditional gender roles. This effect did not occur when naming purported non-traditional couples, however, suggesting that perceived gender roles play a part in gendered mention order biases.

Research from comprehension studies demonstrates the importance of considering the social context that language appears in, including investigations of the
role stereotyping has on processing (e.g., Molinaro et al., 2016). For instance, studies have explored how a listener uses both the semantic information they receive upon hearing a sentence and the social information they have about the speaker (e.g., Tesink et al., 2009). Notably, Van Berkum and colleagues (2008) demonstrated that both of these processes occur very early on in comprehension and are therefore concurrent; who is speaking affects what is being said. In their study, listeners heard sentences that were inconsistent with stereotypical associations of the speaker, for example, a child saying “Every evening I drink some wine before I go to sleep”. They compared this with another task, whereby speakers heard sentences with semantic anomalies, for example “You wash your hands with horse and water” and found that listeners showed the same, early neurological response to both speaker inconsistencies and semantic anomalies – an effect attributable to initial sense-making during comprehension. Thus, a listener immediately takes the social context into consideration when hearing an utterance. This thesis continues the investigation of social context effects, but its focus is on production rather than comprehension. Namely, the studies in this thesis address the extent to which a speaker considers the social context when producing language.

As well as stereotypes, which may be held by individuals regardless of their social group (e.g., Eagly & Steffen, 1984), a person’s own identity and thus their position within a social group may also influence their behaviour. Evidence for this is also found in Hegarty et al.’s (2011) second study, in which men were overall more likely to show a man-first bias whereas women were not, revealing effects of participant gender. This finding highlights the importance of investigating group effects when studying social biases. How people identify according to their identity or group membership are likely to affect how biases occur in language.
To understand this, it is necessary to consider that the notions of belonging and difference are fundamental to social cognition. Research from social psychology has identified stark differences between how people perceive members of their in-group, with whom they share identities, values and other commonalities, verses member of out-groups. Notably, in the 70s, Tajfel and colleagues began identifying how social groups are perceived differently depending on whether a person belongs or does not belong to a group. Social identity theory became a crucial tool for understanding intergroup behaviour (Tajfel, 1978; Tajfel & Turner, 1979), including developing the theory of in-group favouritism. This refers to a common social behaviour to view a group with which one associates more favourably than a group with which they do not; a behaviour that emerges in the first years of life (Donner et al., 2022).

In-group favouritism is considered to occur largely because belonging to a group is an important contributor to a person’s self-esteem or self-worth (Tajfel & Turner, 1986; Brown et al., 1988). An unlimited number of characteristics that are deemed as important to an individual have the potential to affect favouritism, including the sports team a person supports or the type of music they prefer listening to (Ben-Ner et al., 2009). While viewing your in-group favourably is not inherently damaging, this type of favouritism can be linked to prejudice towards out-group members (Aboud, 2003). Extreme examples of this are racism, nationalism, religious extremism, sexuality and gender-based prejudices and other forms of dehumanisation, which manifest in personal interactions and in state policy.

How do intergroup biases manifest in language, and what are the consequences? The Linguistic Category Model, developed by Semin and Fiedler (1988), proposes that there are four distinct categories for talking about other people, and that the categories used depend upon intergroup relations. First, there are so-called descriptive action
verbs, referring to objective descriptions of observable, and neutral, behaviours (e.g., “Kate is hugging Beth”). These are followed by the slightly more abstract interpretive action verbs, referring to a specific action that could be interpreted as positive or negative (e.g., “Kate is comforting Beth”). The next most abstract class is mental or emotional state verbs, which go beyond a specific behaviour and do not have a clear starting or end point (e.g., “Kate feels sorry for Beth”). Lastly, adjectives represent the most abstract class, unattached to a specific behaviour and greatly subjective in nature (e.g., “Kate is empathetic”). The more abstract descriptions imply a perpetual state or personality that will be true across instances outside of the specific named context (if Kate is considered empathetic in the context of comforting Beth, she too will be considered an empathetic person in other situations). This is in comparison to the much more concrete descriptive actions, which are situation-based and temporary (Semin & Fiedler, 1988).

The Linguistic Category Model has been used as a framework for studying dehumanisation, for instance by showing that speakers use fewer mental state verbs when talking about dehumanised people compared to talking about people whose minds they would more spontaneously make inferences about (Harris & Fiske, 2011). It has also been used as a framework for studying agency, for instance in demonstrating that people who are considered as stereotypically agentive, such as men and young people, tend to be mentioned before verbs (e.g., “men vote”) more often than people who are considered as stereotypically non-agentive, such as women and older people, who may be more likely to come before a more abstract noun or an adjective (e.g., “women are voters”, Formanowicz et al., 2017).

In applying the Linguistic Category Model to intergroup descriptions, Maass and colleagues identified differences in how people describe positive and negative
behaviours of members of their in-groups compared to members of out-groups.

According to their theory of a Linguistic Intergroup Bias (Maass et al., 1989; Maass & Russo, 2003), when speaking about in-group members, people are more likely to use abstract adjectives to refer to positive behaviours (e.g., they are kind, warm, empathetic, friendly) and more concrete descriptive actions for negative behaviours (they shouted at, threatened, attacked [other person]). In contrast, when speaking about the behaviours of out-group members, people tend to use concrete descriptive actions for positive behaviours (e.g., they helped, hugged, comforted [other person]) and the more abstract adjectives for negative behaviours (they are confrontational, violent, abusive). In doing so, positive behaviours of in-group members and negative behaviours of out-group members are considered long-lasting, stable and attributable to other situations, while negative behaviours of in-group members and positive behaviours of out-group members are described as being temporary and situation-dependent, rather than a part of the individual’s persona.

When intergroup biases appear in language, they can have real world impacts. One example is that of job interviews. Rubini and Menegatti (2008) studied evaluation reports from hiring committees and found that applicants who had been selected tended to be discussed using positive descriptors from the highly abstract class of the Linguistic Category Model, while their negative descriptors were often of a lower level of abstract, implying that negative traits were observable but temporary properties of the chosen individual. This was particularly evident when selected applicants were coming from the committee’s in-group, raising questions about who is considered a good candidate for a job and who is not.

Interestingly, the Linguistic Category Model and Linguistic Intergroup Bias are not always applicable cross-linguistically. For example, Bonefeld and Beißert (2022)
tested the applicability of the Linguistic Category Model, which was firstly studied in
Italian, to the context of teachers in Germany and implicit biases of ethnicity. Teachers
were asked to describe events involving two children interacting who were either in-
group members (German children) or out-group members (Turkish children). Their
results did not indicate common patterns of linguistic intergroup biases, which would
have been represented by contrasting degrees of abstractness in descriptions of in-group
members and out-group members. Instead, the teachers almost always used concrete
descriptions to refer to both groups. But rather than this indicating that German-
speaking teachers hold no such biases, their results highlighted that the Linguistic
Intergroup Bias may be language-specific. Thus, in developing theories to understand
social biases in language, the social context must always be considered, as biases may
be expressed differently depending on the country, language or other grouping variable.

1.5. Me First and Like Me

In-group/out-group distinctions are not commonly considered by
psycholinguistics in the context of language production. However, research from social
psychology, which has considered social characteristics of the speaker, has shown us
that preferential treatment of an in-group member over an out-group member can
manifest in mention order choices. We find evidence for this in returning to Hegarty and
colleagues’ paper on gender biases in name orders. In Study 6 (Hegarty et al., 2011),
they asked participants to complete three different types of task: writing down the
names of couples in their family, the names of couples from their friendship circle, and
the names of imaginary couples, all heterosexual. When they named imaginary couples,
both female and male participants tended to mention the man first, with male
participants showing a stronger tendency to do so than female participants. However,
when they named familiar couples, effects of participant gender representative of a ‘Me First’ effect (Cooper & Ross, 1975) appeared. On this particular task, own-gender influenced mention order, whereby male participants were more likely to mention the man’s name first and female participants were likely to mention the woman’s name first.

Important findings also come from experimental psychology, with a focus on how conflicts are named (Oeberst & Matschke, 2017). Conflicts provide an interesting case study for group effects, because they can represent very clear in-group/out-group distinctions. In this study, Oeberst and Matschke used Wikipedia to obtain almost 200 titles of historical conflicts between different cultures, referring to the pages in the respective language of the relevant groups to examine how these conflicts had been named (Study 1). In doing so, they found that conflicts were named with the in-group mentioned first and the out-group second in around 85% of cases (e.g., the Polish page would refer to “the Polish–Russian War” whereas the Russian page would do the opposite).

In the same study, but this time in an experimental context, Oebers and Matschke further identified more instances of a Me First pattern of noun ordering, on fictional conflicts involving participants’ in-group country and an out-group country (Study 2). Despite the lower stakes involved in naming a fictitious conflict, participants still showed the tendency to mention their own country first. They also found evidence through using a minimal group paradigm (Study 3) that distinctions of in-groups and out-groups can even be entirely arbitrary and still have an effect on mention order. Here, participants were assigned to fictitious genetic groups before naming a conflict between their group and another. Again, participants opted to name their in-group first,
suggesting long-term association with a social group is not required for this mention order bias.

Oeberts and Matschke have shown how a Me First effect can occur regardless of the importance of the event (i.e., a real war versus a fake war), or the concreteness of the social group (real groups versus fake group). It seems likely that there may then be other relevant ways in which social biases modulate production. In particular, people might be affected by their relative social similarity to a participant in an action. If so, this would reflect what is known as the Like Me hypothesis of social cognition. This states that seeking out similarities between oneself and someone else is a fundamental social process, beginning as early as when an infant recognises similar motor information, such as bodily movements and facial expressions, in other people (Meltzoff, 2007). In recognising self-other equivalences, we may conclude that we share similar cognitive abilities and emotions. Thus, when a speaker is describing an interaction between two people, it seems possible that part of conceptualising this event would involve identifying potential self-other equivalences, and that who they mention first might reflect who they are more socially similar to.

Some evidence for the effect of relative social similarity comes from one particularly notable psycholinguistic paper by Iliev and Smirnova (2016). Their series of studies attempted to empirically validate the Me First hypothesis, i.e., that the first entity a speaker mentions is that which is more closely associated with them (Cooper & Ross, 1975). They did so with three different approaches. Their first study addressed four categories: brand preferences of consumers, political affiliation, religious beliefs, and nationality. They created word pairs for each of these categories using nouns that represented opposition within the category (for example, Honda and Toyota, democrat and republican, Christian and Muslim, USA and Canada), and tested how prevalent
each pair was on websites and forums relating to these topics (e.g., Honda forums and Toyota forums, liberal websites and conservative websites, Christian websites and Muslim websites, and websites with URLs ending in .us or .ca). Results showed a strong tendency, across all four categories, for the mention order of noun pairs to reflect the alignment of the website. For instance, the order “Honda and Toyota” was more common on Honda forums than forums of its brand competitor, and the order “Christian and Muslim” was more common on Christian websites than on Muslim websites.

In their second study, Iliev and Smirnova tested gendered word orders (e.g., “sister and brother”, “father and mother”) in written text from corpus data, firstly looking at literature and then blog posts. They found that male lexical items were mentioned first overwhelmingly more frequently than female lexical items, for both male and female authors, although female authors tended to show a slightly lower man-first bias. They demonstrated this man-first bias again in an experimental study, for which they asked participants to answer the question “what are the two genders?” On the whole, participants mentioned male first and female second, although women were less likely to do so than men. In this study, they also tested university affiliation (“What are the two top-ranking universities in the Chicago area?”), political stance (“What are the two main political parties in the US?”), and race (“What are the seven continents?”). Their results showed that people were more likely to mention their own university and the political party they most aligned with first. For investigations of race, they predicted that the first continent mentioned would be the one most closely related to their race. And, indeed, they found that white people were more likely than non-white people to mention North America first and Asian people were more likely than non-Asians to mention Asia first. Other racial groups showed similar trends, but had extremely small participant numbers.
Iliev and Smirnova’s (2016) studies demonstrate how being more socially similar to one group over another can affect who a speaker mentions first when producing binomials, but what about more complex language? Interestingly, in their studies, they also varied the type of binomial, comparing “and” conjunctions (e.g., “USA and Canada”), “or” conjunctions (e.g., “USA or Canada”) and no conjunctions (e.g., “USA, Canada”), and found similar trends for each type. This finding suggests that serial mention order alone may be the important factor of first mention biases, over (for instance) distinctions of grammar.

1.6. Mention Order and Comprehension

The research I have just reviewed points towards the assertion that language production contains social biases that either reflect and maintain conventional social hierarchies (for instance, a male-oriented society) or reflect group identification (such as Christians tending to mention Christian first and Muslim second). With the acknowledgement that social biases exist, in language and beyond, it is important to then consider what the effects of these biases might be. Regarding language, these effects reasonably relate to comprehension and semantics.

A large area of psycholinguistics has been dedicated to understanding how people make semantic inferences from grammatical distinctions, which includes the order in which nouns are mentioned. According to MacWhinney (1977), the focus of a sentence often falls on the sentence subject, and so whichever entity is afforded this position will act as an anchor for both speaker and listener (Kelly et al., 1986). Due to their relative subtlety, the effects of mention order are perhaps easier to overlook, compared to more overt biases such the Linguistic Intergroup Bias or stereotypical
descriptions of people and behaviours. Nonetheless, they may have just as serious implications.

Prior research has demonstrated that semantic inferences are made from sentence subjecthood and thematic role. In fact, deriving meaning from syntax is identifiable as early as 2 years old (Arunachalam & Waxman, 2010). One of the earliest papers, by Johnson (1967), considered this in the context of determining animacy from subject position. When participants encountered a sentence with nonsense nouns, such as “The NIJ hurt the GAQ”, the first-mentioned noun, both the subject and the thematic agent, were assumed as more animate than the second-mentioned noun, which is both the direct object and thematic patient.

An important study by Gleitman and colleagues (1996) also shows that even when these thematic roles are less clear, for instance when both entities involved in an interaction are executing the same action together, comprehenders still infer meaning from mention order. They tested this using symmetrical predicates, which are verbs that are supposed to imply symmetry between two nouns (such as meet and match). Participants were asked to rate the similarity of two sentences in which these predicates were embedded, and which were syntactically identical other than the order of mention, for example “Meryl Streep met my sister” with “my sister met Meryl Streep”. Ratings were given on a 5-point Likert scale from “do mean the same” to "do not mean the same”. Despite the only difference between the sentences being the order of nouns, ratings indicated that participants construed the sentences as having different meanings.

In psycholinguistics, semantic inferences from mention order have been important for studies of language development and interpretation. In social psychology, however, semantic inferences have been discussed in the context of societal
implications and biased representations. This includes discussions of the way in which the passive voice is used to talk about social issues.

The passive voice is often avoided by academics, for the purpose of increasing the readability of written work (though see Ferreira, 2021 for a pro-passive standpoint). Aside from readability, the passive voice has come under fire for the ways in which it has been and is used to talk about victimhood and culpability, particularly in descriptions of crimes. Or, as Alexis Pauline Gumbs has put it:

*The problem with the passive past tense is that it obscures the relationship between subjects and action, between what we do and how it impacts other people... Making action invisible, not only forecloses robust accountability for people who harm other people, it also disempowers us all by hiding the myriad actions we could take to end cycles of violence. If someone did something, we can do something else, we can do something now.*

- Gumbs (2018)

Research from the domain of victim perception demonstrates how mention order can lead to stereotypical perceptions of agency and responsibility in descriptions of violence. For instance, Bohner (2001) investigated the uses and effects of the passive voice in relation to sexual violence, conducting a study in which students viewed a rape scene and gave a written description of what they had viewed. They also provided judgements of responsibility for both the abuser and the victim and provided information about their own attitudes towards sexual violence via a rape-myth scale. Results showed that participants often used the passive voice to talk about the rape, mentioning the victim first and the abuser second (e.g., “the woman was attacked by the
man”), which Bohner argues results in placing more focus on the victim while the abuser is afforded a background role.

This is not necessarily an inherently bad outcome – there is an argument to be made about how focusing on someone who has experienced sexual violence is an important part of humanising and empathising with them, whereas focusing on the abuser runs the risk of further removing agency from the victim. However, through the collections of additional measures, Bohner (2001) also found that participants who were more likely to accept rape myths (e.g., that rapes occur when the victim is wearing particular clothes) and who had used the passive voice to describe the event tended to place more responsibility on the victim and less on the rapist. Thus, mention order in this context has an important effect on the semantics of violence. In a society where victim-blaming, such as that which occurs due to rape myths, is prevalent, biases of first mention may present a worrying influence on comprehension.

Transitive sentences, both active and passive, reasonably inspire semantic inferences from mention order due to distinctions of grammatical and thematic roles between nouns. Interestingly, there is also some evidence to suggest that readers will infer meaning from mention order without these distinction, for instance in binomials or in conjoined sentences where both nouns share the subject position.

Oeberst and Matschke’s (2017) work on conflict naming demonstrates this in the case of binomials. How a conflict is named can be an important indicator of group membership and intergroup identity, but they may also reveal semantic biases that could influence a listener or reader’s interpretations. Conflicts, including wars between nations, are sadly a regular occurrence, and so literature and the media present us with conflict names on a regular basis. For instance, current news reports around the globe have devoted attention to Ukraine since Russia invaded the country in February 2022,
including UK publication The Guardian’s daily report which they have so far always titled “Russia-Ukraine war: what we know on day [e.g., 67] of the invasion”. But what are the effects of hearing either “The Russia-Ukraine war” or “The Ukraine-Russia war” for someone who is part of neither group? Does first mention indicate the initiator of the conflict, the likely “winner” of a war, or maybe who the messenger has denoted should receive the most sympathy or attention? Indeed, in Studies 4 and 5, Oeberst and Matschke tested the semantic effects of first mention on conflict titles and found that groups were often perceived as being of a higher importance or power when they were mentioned first rather than those that were mentioned second.

Further evidence for the effect of mention order in binomials on semantic inferences comes from the work by Hegarty et al. (2011) on couple names. In Study 4, they tested the effects of mention order on gender stereotypes, this time looking at names of same-gender couples. Participants were asked to write the names down of imaginary same-gender couples, and then to provide information on how each partner differed from one another. Results showed that the first-mentioned name was often attributed to more supposed “masculine” traits than the second-mentioned name, for instance in being bigger and stronger. This suggests that gender biases in mention order may be reflective more of perceived gender traits rather than a fixed gender binary of man-first and woman-second.

Mention order may signify other properties about people in a description, namely relevance. Evidence for this comes from Kesebir (2017), who presented participants with a text including sentences where a mother and father character were involved in a description of an event (Study 6a), counterbalancing which parent appeared first or second (e.g., My [mother and father] / [father and mother] have met the coach last week… My [mother and father] / [father and mother] told me afterwards
that they want me to work on my weaknesses, but not worry too much about them).

Participants then answered a question on perceived relevance of the mother and father characters, i.e., “who is more involved in this student’s tennis life?” When the father was mentioned first, participants rated them as more involved than the mother, but when the mother was mentioned first, participants did not show a statistical difference in rating the mother as more involved than the father. Thus, first mention was found to not only influence semantic interpretations of the importance of the individuals in the sentence, but these interpretations were also modulated by effects of gender that highlighted imbalances between mother and father and their parenting expectations.

The context of the event and the identity of the reader may also modulate interpretations of mention order. In an additional experiment (Study 6b), Kesebir presented a text about protest, using the nouns women and men – for example, Some of the town’s [women and men] / [men and women] are out on the streets, talking to the locals. Participants then answered the question “which group is playing a more central role in the organized protests?”, choosing between the women and the men. When the women were mentioned first in the text, participants more frequently chose the women as more central than the men, and when the men were mentioned first, participants more often chose the men as more central than the women. Male participants were also significantly more likely than female participants to claim the men as more central. From this, Kesebir suggests that people will not only perceive a first-mentioned lexically gendered noun as relatively more relevant than a second-mentioned noun, but also that the way in which the context is gendered, comparing a parents’ evening or a discussion with a team coach with a protest, will also affect how first mention influences perceptions. Crucially, these semantic inferences occurred without distinctions of grammatical or thematic role between the nouns.
1.7. Thesis studies

Research on socially biased mention order can be found in the domains of psycholinguistics, social cognition and social psychology, however these domains do not always meaningfully refer to each other in their investigations. The studies in this thesis also seek to explore social biases of mention order, bringing these domains together.

In Chapter 2 I will investigate whether speakers demonstrate mention order biases when describing interactions between people who either differ in gender or differ in race. Gender and race were chosen for two reasons: first, gender has already received some focus in the study of socially biased mention order but demands to be explored further. Second, these are two highly salient and influential social constructs that have had important consequences for society, but there are also limitations in talking about one without the other considering the ways in which gender and race often interlock (Taylor, 2017; Crenshaw, 1986). This is, of course, true for multiple categories of social group, however by starting with gender and race I hope to develop a paradigm that can be used for other types of social bias.

In Experiments 1 and 2, I aim to investigate if who a speaker mentions first is influenced by societal hierarchies (e.g., if people are more likely to mention a man first and a woman second due to androcentrism, or a white person first and a Black person second due to racial hierarchies) or by a Like Me bias in which they mention the person of their own gender or of their own race first, over a person of a different gender or race. This study therefore provides a novel way of addressing socially biased mention orders, in that it considers both the speaker’s social context and their own social characteristics. I choose to label this a “Like Me bias” rather than a “Me First bias”, because I believe
the former makes more room for the notion of relative social similarity. I would argue that “Me First” suggests the speaker sees themselves as affiliated with one referent and not the other (i.e., a Polish speaker describing a conflict as the Polish-Russian war does not identify as Russian), or that their mention order bias reflects self-prototypicality (e.g., mentioning entities that are central to the speaker first when ordering nouns, such as “man and machine” (Cooper & Ross, 1975). “Like Me”, which has roots in social cognition, appreciates that the speaker reasonably shares similarities with both individuals to be described, but the framing indicates that there exists some relative social similarity between the speaker and both the person they mention first (i.e., that they are more like the speaker) and the person they mention second (i.e., that they are less like the speaker).

Black and white participants viewed images of interactions depicting either a woman and a man of the same race interacting, or a Black and a white person of the same gender interacting, and I measured who the participants mentioned first. In these studies, I also investigate the extent to which syntax influences how mention order biases occur in production, comparing descriptions made by participants using transitive syntax (e.g., “Beth is hugging Kate”) and conjoined syntax (e.g., “Beth and Kate are hugging each other”). In Experiment 3, I explore the extent to which the Like Me bias is influenced by perceptions of agency conjured by visual perceptions of the people depicted in these events. And lastly, in Experiment 4, I return to effects of syntax, this time comparing transitive active sentences with passive sentences (e.g., “Beth was hugged by Kate”), to determine the effect of thematic role on mention order choices.

In Chapter 3, Experiment 5 investigates the implications of mention order biases on comprehension, asking to what extent does first mention affect semantic inferences. Here, I focus on an important property of person perception, testing judgements of
agency. To investigate this, participants read short sentences describing an interaction in which one person was mentioned first and the other second and provided ratings of agency for the two people. I measured who participants rated as the more likely agent, predicting that this would be the person who was mentioned first.

This experiment also tested effects of gender, recognising the established effect that gender has on agency perceptions whereby men are more often associated with the properties of an agent than women (e.g., Eagly & Johannesen-Schmidt, 2001), and the emerging evidence showing that mention order has an effect on gendered perceptions of agency as it relates to relevance and importance in a relationship context (e.g., Kesebir, 2017). Thus, Experiment 5 also aimed to determine if people are more likely to perceive a person who is mentioned first as agentive when that person is a man rather than a woman. I consider that such a bias could be population general (i.e., as in Kesebir) or that it could be modulated by a Like Me effect to view the first-mentioned person as more agentive when that person is the same gender as the reader rather than a different gender. Overall, Chapter 3 seeks to demonstrate how mention order is not passive or even necessarily harmless, and that it can instead have serious implications for the ways in which social and cultural narratives are structured.

Lastly, Chapter 4 seeks to explore another important area of social cognition, the perception of animacy, in the context of language production. The two studies in this chapter apply the influential theory of conceptual accessibility and the related animacy bias to descriptions of interactions between humans and robots. In Experiment 6, participants viewed and described images of human-robot dyadic interactions and I measured who they mentioned first – human, or robot. Humans are higher on the conceptual accessibility hierarchy than robots, and so I predicted that participants would show a robust human-first bias. Following this, in Experiment 7 I tested who
participants would mention first if the inherent accessibility of the human and robot
groups was altered. Using a narrative, I manipulated perceptions of animacy for these
groups, presenting the robots as anthropomorphised and thus “more similar” to
participants by giving them human-like characteristics and the humans as “less similar”
to participants via dehumanisation. Narrative was chosen as a methodology because of
its power to suspend belief. Additionally, previous research has shown that fictional
contexts can have immediate and influential effects on processing, to the extent that
they eliminate what would otherwise be pragmatic anomalies (Nieuwland & Van
Berkum, 2006; Filik & Leuthold, 2008), such as “the robot is marrying the human”.

I predicted that this manipulation would increase the likelihood that participants
would mention the robot first, not the human. In doing so, I investigated the extent to
which perceptions of mind and self-other equivalences determine who a speaker is
likely to mention first in their descriptions. Consequently, Chapter 4 builds on the work
in Chapter 2, by exploring the Like Me bias of language production. Importantly, these
studies offer important insight into how humanness is perceived and defined.

1.8. Summary

To summarise, previous work from psycholinguistics has shown how mention
order is influenced by conceptual factors, whereby a speaker describing an event will
most likely mention the most conceptually accessible entity first, for instance by
mentioning a more animate entity before a less animate entity. However, the current
knowledge in the field is not equipped to make predictions of mention order when the
event the speaker is describing involves two people. Similarly, research from domains
of social cognition and social psychology have addressed how social biases manifest in
language, including who a speaker is likely to mention first, but this research has not so
far address event descriptions and does not often consider the importance of syntax. Furthermore, research of socially biased language often focuses on gender, meaning other social constructs that may have an independent or intersectional effect, such as race, have not been addressed at all in this context. Both gender and race have huge impacts on society, and so I have chosen to study them together, appreciating that gender and race create their own distinct social groupings, but that they also interlock.

Importantly, who a speaker mentions first has been found to influence semantic inferences, with first mention sometimes associated with greater importance, power or relevance, and even stereotypical gender traits. Again, the research that has addressed this is few and far between, despite the important implications of these semantic biases on wider society.

In bringing together theories and paradigms from different research domains, I aim to provide a fuller understanding of how social factors influence mention order, both in production and comprehension, and demonstrate why research of society and language must always take a multidisciplinary approach. This thesis ultimately seeks to show how mention order can both indicate and reinforce a speaker’s social biases, as well as the hierarchies in our society that demand dismantling.
2. Study 1, Experiment 1 – 4: Effects of gender and race on mention order and perspective

2.1. Introduction

Psycholinguistics has long been engaged in studying how speakers order words in utterances, with the overarching notion that mention order is not arbitrary or random, and instead may be driven by various perceptual and cognitive factors. When describing interactions or associations between objects, mention order is ever-present, with the speaker having to make rapid decisions about what to talk about first and how best to represent the event. This often involves decisions of whose perspective to take – in other words, from whose eyes to describe the event. Mention order therefore provides language researchers with a rich lens through which to study a range of issues, from how a speaker develops production strategies to how biases in our social world manifest in language.

Numerous studies, addressing both child and adults speakers and speakers of diverse languages, have extensively researched differences in properties of first-mentioned entities over second-mentioned entities, for example, identifying the tendency to mention animate objects first and inanimates second (e.g., Branigan, Pickering, & Tanaka, 2008; Lempert, 1990; Lake, Cardy, & Humphreys, 2010; Tanaka et al., 2011; Øvrelid, 2004; Prat-Sala & Branigan, 2000; Ferreira, 1994, Christianson & Ferreira, 2005; Franz et al., 2021). Research has also detected biases of mention order in noun pairs joined by a conjunction, reflecting conventional social hierarchies and stereotypes. For example, there is an argument that the way in which noun pairs of lexical gender (nouns relating to maleness or femaleness, such as “husband and wife”) show a man-first bias, reveals preferences for a male-oriented or androcentric world
view (Hsiao et al., 2021) and stereotypes of gender roles (Hegarty et al., 2011; Hegarty et al., 2016).

There is some evidence to suggest that mention order may be dependent on characteristics of the speaker. Cooper and Ross (1975) first described a “Me First” principle of ordering words in conjunctions, whereby the first mentioned word is often the one that refers more to properties of the speaker; a theory that was later tested and supported by Iliev and Smirnova (2016). Additionally, who the speaker mentions first may be influenced by which person is closest in relationship to the speaker (Hegarty et al., 2011; Tachihara, Pitcher, & Goldberg, 2019; Tachihara & Goldberg, 2020), which person is from the speaker’s in-group (Oeberst & Matschke, 2017) or which person appears more relevant to the speaker in that context (Kesebir, 2017). Word order can therefore indicate properties about the speaker, including their relationships and their attitudes about the world.

We propose that the existence of patterns like these should motivate psycholinguistics to consider how identity and society might lead to biases in language production. Our research seeks to take a multidisciplinary perspective, incorporating previous research that has considered the identity of the speaker when predicting mention order biases with psycholinguistic frameworks that have focused on determinants of syntactic choices in language production.

Specifically, we are interested in how a speaker’s gender or race may influence their mention order choices. Gender and race are examples of two highly impactful social constructs that not only affect how an individual exists within and sees the world but have large effects on a speaker’s language. Biases in mention order, as related to these factors, may have the potential to inform us about social cognition, as well as revealing how hierarchies in society are sustained through language.
2.2. Background

Between disciplines, researchers have explored different factors associated with mention order biases. This has included the role of animacy (e.g., Branigan et al., 2008), conventional social hierarchies (e.g., Hegarty et al., 2011; Hsiao et al., 2021), schematic representations, relevance to the speaker as a social being (e.g., Kesebir, 2017), personal relationships (e.g., Tachihara & Goldberg, 2020), and in-group and out-group distinctions (e.g., Oeberst & Matschke, 2017). Work regarding these factors contributes to our understanding of mention order as systematic.

In the 80s, Kathryn Bock established a crucial explanation for why speakers show patterns of mentioning some words before others, which she termed conceptual accessibility, or the ease with which the mental representation of some potential referent can be activated in or retrieved from memory (Bock; 1987; Bock, 1986a, 1986b; Bock & Warren, 1985; McDonald et al., 1993). Bock proposed that speakers tend to construct utterances based on which representations are easier to retrieve from memory, and are therefore easier to bring to mind, or conceptualise. Doing so is considered beneficial for the efficiency of fluent utterances, as beginning with the more easily retrievable entity takes some pressure off the speaker’s working memory, which can instead be directed towards retrieving the less accessible entity later on. As a result, more conceptually accessible entities are often expressed earlier in utterances than less conceptually accessible entities.

Conceptual accessibility offers a convincing explanation for some of the mention order biases we find in production. For instance, animate entities are considered more conceptually accessible than non-animate entities due to their significance to knowledge networks and therefore their importance in human thought.
Numerous studies have shown that, as a result, speakers generally mention a more animate object before a less animate object when describing an event or scene involving the two, for example saying “the researcher was near the cat” rather than the opposite (Bock 1982, Bock 1987; Bock & Warren 1985; Levelt 1989; Bock & Levelt 1994; McDonald et al., 1993; Ferreira 1994; Christianson & Ferreira 2005; Tomlin 1995; Downing and Noonan 1995).

Interestingly, speakers tend to do this across different syntactic structures, including both active and passive sentences (Prat-Sala & Branigan, 2000), suggesting the order of mention and subject attribution can be more important to speakers than thematic role when talking about objects of different levels of animacy.

Thus, the ordering effect brought on by animacy distinctions often results in a speaker mentioning the most animate entity first, even when this means it will not be the event agent. The animacy bias has been thoroughly studied (see Branigan et al., 2008 for review), with research discovering this behavioural pattern cross-linguistically (e.g., Prat-Sala et al., 2000; Tanaka et al., 2011; see also de Swart et al., 2008), in neurotypical and neurodiverse populations (Lake et al., 2010) and in young children (Buckle, 2020; Franz et al., 2021).

There is also some evidence that mention order is influenced by conventional social hierarchies. For instance, Hegarty et al. (2011) have found that, in the naming of heterosexual pairs or couples, a man’s name is mentioned first more often than a woman’s name (Studies 1, 2 and 3), and stereotypical traits of masculinity tend to be associated with a first-mentioned name over a second-mentioned name when providing information about same-gender couples (Study 4).

Additional research on stereotyping and mention order has linked the first-mention bias to schematic representations. This refers to the ways in which a person’s
knowledge about the world and their cultural and social background influence how they conceptualise events, guide behaviour and think about objects and concepts (Fiske & Taylor, 1991). Namely, Kesebir (2017) has argued that it may be the relevance, or centrality in context, of a concept to a speaker that will increase its accessibility and therefore the likelihood that the speaker will mention it first. Using gender as an example, she proposes that in the context of fashion, a speaker might be more likely to say best-dressed actresses and actors on the red carpet over best-dressed actors and actresses, ostensibly due to the feminine stereotype that fashion holds. The order in which lexical gender is presented in conjoined noun phrases would therefore reflect which gender in that specific context is more relevant (Kesebir, 2017). Context, however, is still ultimately affected by an individual’s schema. If a male fashion designer who frequently dresses male actors and ascribes considerable importance to men’s fashion over women’s was talking about the red carpet, we might expect that he would mention actors first and actresses second.

Relevance to the speaker as an individual also involves interpersonal factors, such as the relationship between the speakers and the people they are talking about. In the domain of cognitive linguistics, Tachihara and Goldberg (2020) studied the ordering of proper names, specifically addressing how a speaker talks about couples. English and Japanese speakers were asked to name three important couples in their lives, and the order in which they mentioned the individual names was recorded. They found that that speakers tended to mention the name of the person they felt closest to.

This study follows another by Hegarty et al. (2011, Study 5), who collected Christmas cards addressed to couples and looked at the order in which the names were written. They found that the person mentioned first was often the person closest in relationship to the sender, which tended to correlate with the tendency for the first-
mentioned person to be the same gender as the sender. They showed this ‘Me First’ effect (Cooper & Ross, 1975) again in Study 6, when participants wrote down the names of heterosexual couples they knew personally, with results showing that male participants were more likely to mention the man’s name first and female participants were more likely to mention the woman’s name first. When the couples were familiar to the speaker, the man-first effect of gender stereotypes as found in Studies 1 – 3 appeared to be replaced by an own-gender effect. The authors attributed these findings to an effect of familiarity and of closeness to the speaker, considering that women tend to form closer relationships with other women, while men tend to form closer relationships with other men. However, they did not collect data on how gender informed participants’ personal relationships, so this explanation may not cover the full story.

It may also be possible that people are likely to show a Me First effect because of perceived similarity linked to group identity, rather than interpersonal closeness. Evidence for this comes from Iliev and Smirnova (2016), who found that mention orders of conjunctions (e.g., “liberal and conservative”) demonstrates group association (for instance, they found that this mention order is more likely to occur on Democratic websites rather than Republican websites). This would suggest that it could be possible to predict mention order based on social factors at the group level, rather than at the individual level.

Oeberst and Matschke (2017) provide additional evidence for a Me First effect on first mention, investigating the association between mention order and conflict. They found that speakers will often prioritise names of groups to which they belong over out-groups, thereby mentioning their ingroup first. For instance, a Polish speaker may show a tendency to say “the Polish-Russian war” whereas a Russian speaker may say “the
Russian-Polish war”. Their work highlights the importance of group identity and social similarity, as members of in-groups typically share similar properties (e.g., beliefs, culture, or appearance) which differentiate them from members of an out-group, on mention order choices. Thus, there is some evidence that characteristics of the speaker can be an influencing factor of mention order decisions, at both an individual level, regarding relationship closeness or perhaps familiarity, and at a group level, regarding similarity associations with an ingroup over an outgroup.

Noun pairs provide us with an interesting foundation of the effect of social factors on mention order. Our study seeks to expand on the previous research, by investigating if socially driven mention order biases exist in more complex language, namely that which is used to talk about interactions between people. We aimed to investigate how social factors, which make a speaker more socially similar to one person and less socially similar to another, affect who the speaker mentions first. Socially biased mention order is tied to perspective taking, or whose eyes an event is viewed through. Consequently, mention order is an important social issue, as biases of perspective have implications for semantic inferences.

Furthermore, while some research has considered how the speaker’s gender influences their mention order choices (e.g., Hegarty et al., 2011), as to our knowledge, effects of the speaker’s identity on social language in a racial context has not yet received much attention. Iliev and Smirnova (2016) addressed this to some extent, measuring if a person’s race influenced whether they would mention a continent more associated with their race than another continent, and indeed found that white people were more likely to say “Europe and Asia” than “Asia and Europe”, and Asian people showed the opposite patterns in their mention order. However, their study conflates race with nationality and continents are not so simplistically divided by associations with
just one race. Thus, research on mention order still has a lot to learn about the effects of racial identity.

Like gender, race is another social construct that has been used to establish conventional social hierarchies and can be a poignant marker of ingroup and outgroup distinctions. We therefore chose to investigate the effects of gender and race, because these are two highly salient factors of social hierarchy and of social groups. Regarding social hierarchy, we consider effects of gender as it relates to a male-oriented world view (or androcentrism), and of race as it relates to effects of the white default, whereby whiteness as the dominant racialised group in the context of this research positions all others as marginalised. Regarding social identity, we consider that gender and race can be determinants of whether someone identifies more or less with a social group; namely, with people of their own gender versus with people of a different gender, or with people of their own race versus with people of a different race.

Gender and race have important effects on cognition. Outside of linguistics, research has found that people tend to be better at remembering and recognising the faces of people of their own race (Meissner & Brigham, 2001; Vingilis-Jaremko et al., 2020). By 9 months, infants begin to show a better ability at recognising own-race faces than other-race faces (Kelly et al., 2007) and by preschool age, young children are already making social categorisations based off race (e.g., Dunham et al., 2013; Perszyk et al., 2019) and gender (Rawan et al., 2020; Thompson, 1975). People often align or associate with others according to gender and race – for example, forming closer relationships with people of their own gender (Caldwell & Peplau, 1982) and looking to receive support from people of their own race than from people of a different race (Davis & High, 2019), or experiencing a stronger negative response on witnessing instances of ostracism targeted towards members of their own race than towards
someone of a different race (He et al., 2021). The strength of ingroup associations will differ from person to person, with gender and race existing as important facets of some people’s identity and not of others. It is currently unclear if the degree to which identifying with one’s own gender or race modulates the likelihood of these social factors affecting a speaker’s mention order biases.

In the current study, we sought to understand if a speaker’s mention order choices will differ depending on their gender and their race, when producing utterances where both factors are observable features of the event. To do so, we used a picture-description task to measure mention order when describing interactions between two people who either differed in gender or in race, analysing who participants chose to mention first. To test the effect of social similarity, participants in this study were also Black and white women and men. Thus, participants viewed images in which one figure was always in the participant’s gender in-group and the other was in their gender out-group, and images in which one figure was always in the participant’s racial in-group and the other was in their racial out-group.

Descriptions were formed using symmetrical predicates (Gleitman et al., 1996), which are verbs that imply a symmetry of action between the two figures because both figures are executing the same action (e.g. meeting, hugging, touching). However, when used in a sentence, they require the speaker to make a decision as to which figure should be mentioned first (e.g., Jane met Fred vs Fred met Jane). Images were presented as graphics interchange format files (GIFs), switching rapidly on their axis, to avoid any potentially confounding left-right order effects on participants’ descriptions (Maass & Russo, 2003).

Participants learnt names of the individuals presented in the interactions during a memory task, before undertaking the language task. Using names allowed us to avoid
confounding effects that may arise from common patterns of lexical gender or frozen noun pairs (Cooper & Ross, 1975; Mollin, 2012), especially in conjoined noun phrases such as “man and woman”. Doing so also helped avoid effects associated with political correctness and “colour blindness” that might arise from distinguishing between two figures based of the same gender with or without referring to their race (Norton et al., 2006).¹ This approach also circumvents potential ambiguity of descriptions and avoids highlighting the gender or racial elements of the experiment.

The design of this study therefore provided participants with the reasonable chance to make either figure the sentence subject. Analyses measured if participants routinely prioritised one figure over another, i.e., by mentioning them first in their event descriptions and therefore taking their perspective. We assessed how this tendency varied based not just on the identity of the figure, but also on the identity of the participant.

We considered the possibility that biases learnt from being exposed to hierarchies in society may influence mention order. Regarding gender, we might expect participants to mention men first over women due to androcentrism. In a similar vein, we might expect participants to mention white people first over Black people because we are testing populations from white-majority countries where whiteness is considered the default. We tested this effect of social hierarchies by measuring the proportion of

¹ Results from a comparable picture-description task in a pilot study emphasised the importance of giving speakers an unambiguous way to refer to people of different races during picture description tasks. We found that 47% of Black participants would use racial descriptors in their utterances, such as “the Black woman” or “the white man”, whereas only 10% of white participants did so, instead opting for descriptors of hair type, clothing, facial expression and so on, often resulting in longer utterances. Using names for figures therefore leaves the participant with only the task of describing the interaction, rather than thinking about how to appropriately reference the people and avoids possible production differences between our Black and white participants that could confound effects modulating mention order biases.
gender trials showing a man-first bias, and the proportion of race trials showing a white-first bias.

Additionally, we predicted that participants would demonstrate a tendency to take the perspective of the person most like themselves, meaning the person who shares either the same gender or the same racial identity as the participant, and therefore that they would mention this person first, thereby taking their perspective. We call this effect the ‘Like Me bias’. For example, we predicted that male participants would show a Like Me gender bias to mention the man before the woman on gender trials whereas female participants would mention the woman before the man, and that Black participants would show a Like Me race bias to mention the Black person before the white person on race trials whereas white participants would mention the white person before the Black person. Thus, we investigated the extent to which mention order, when describing an interaction between two people, was influenced by relative social similarity between the dyads and the speaker.

Furthermore, we took an intersectional approach by exploring the interaction between race and gender on first mention, investigating if participants would be more likely to mention the person matching their own gender first if that person is also of the same race as themselves, or if they would be more likely to mention the person of the same race as themselves if they are also the same gender as the participant (for example, a Black female participant showing a stronger Like Me race bias when describing an interaction between a Black woman and a white woman, than when describing an interaction between a Black man and a white man).

We therefore conducted two related but distinct analyses; firstly, looking at the role of conventional social hierarchies (man-first and white-first) and secondly, testing our Like Me hypothesis (a Like Me gender bias to mention the person of their own
gender first, and a Like Me race bias to mention the person of their own race first). Attempting to study social effects on mention order through only one possible lens of social cognition may result in failing to consider the fuller picture. Measuring the effects of gender and race on mention order according to conventional social hierarchies only would miss potential effects of group identification, whereas doing so by only considering the Like Me hypothesis would miss potential effects of social order. Importantly, these two effects may not be mutually exclusive, as shown by Iliev and Smirnova (2016), who found a man-first effect in lexical gender conjunctions (e.g., “male and female”), but whose results also showed that women were significantly less likely than men to demonstrate this bias.

We were also interested in how a Like Me bias might manifest depending on syntactic characteristics of speakers’ utterances. Unlike the previous work which has focused on binomials, the current study tests mention order biases in utterances requiring a designation of grammatical function and thematic role. To do so, we examined the extent to which syntax would affect mention order, examining the interplay between grammar and social biases. Thus, we crossed dyad conditions (different-gender dyads and different-race dyads) with a manipulation of sentence structure.

Specifically, we compared active transitive sentences, which designate clear subject/object positions and thematic roles of agent and patient at first and second position respectively (e.g., “Kate is touching Fred”), and conjoined sentences with NP conjunctions where grammatical function does not differ between the figures (e.g., “Kate and Fred are touching each other”). Evidence that there might be a difference between these two sentence types in how mention order biases manifest comes from Tanaka et al. (2011), who compared noun phrase conjunctions with transitive sentences.

60
for the presence of an animacy bias. As expected, they found that speakers were highly likely to produce an animacy bias on transitive sentences, however speakers were not reliably likely to mention the more animate entity first for noun phrase conjunctions. Additionally, when testing the effects of imageability, as a property of conceptual accessibility, on mention order, Bock and Warren only found effects when participants were producing sentences with variations in grammatical function between the nouns, such as transitive sentences, and not for conjoined phrases (Bock & Warren, 1985). These finding suggests that the syntactic differences between transitive sentences and conjunctions (that the former involves distinctions of grammatical function while the latter does not), may result in different expressions of mention order (see also Branigan et al., 2008 for a review).

In Experiments 1 and 2, speakers produced transitive active sentences and conjoined sentences, and we measured whether effects of accessibility resulting in mention order biases would be modulated by grammatical function distinctions. We predicted that there would be a stronger effect on transitive trials, i.e., when the first mentioned person was the sentence subject and the second mentioned person was the direct object, than on conjoined trials where both people constituted the sentence subject.

In Experiment 3, we considered why speakers might mention someone who is more socially similar to themselves first. One explanation for the Like Me bias is that speakers perceive people from their in-group as more agentive than people from their out-group. To test this, we measured perceptions of agency, to determine if people show a Like Me bias to perceive a person of the same gender as themselves as more agentive than a person of a different gender, and to perceive a person of the same race as themselves as more agentive than a person of a different race.
Then, in Experiment 4, we explored the role of agency further, by testing how mention order biases are affected by distinctions of thematic agent and patient. This time, participants produced transitive active and transitive passive sentences (e.g., “Kate is being touched by Fred”) and we investigated the extent to which the Like Me bias is affected by thematic role.

Finally, we conducted some exploratory analyses to understand more about what might motivate mention order biases. We examined how individual differences (identification with their ingroup and social attitudes) might affect mention order biases. We predicted that participants would be more likely to demonstrate a Like Me bias if they showed stronger identification with their in-group, scoring higher on group-level self-investment and self-definition (Leach et al., 2008, Experiments 1 and 4). Secondly, we conducted a control analysis on race trials, which depicted interracial and same-gender relationships, exploring the degree to which attitudes towards these factors affected and the Like Me race bias (Experiment 2). Findings from Experiment 1 prompted us to explore two additional effects more formally in Experiment 2 and beyond. First, we considered how the valence of the event may have modulated who participants mentioned first, categorising the interactions as either intimate, casual or negative events. Second, we considered how memory for names may have influenced mention order, exploring if participants were influenced to mention certain people first because they had memorised their names more successfully.

In sum, mention order when describing dyadic interactions may be influenced by at least two factors: conventional social hierarchies or a Like Me bias. We predict that, if social hierarchies contributing to an androcentric or white-defaulted world view affect mention order, participants will be more likely to mention a man first than a woman when describing woman/man dyadic interactions, and a white person first than a
Black person when describing Black person/white person dyadic interactions respectively. If people show a Like Me bias, which results from relative similarity between the speaker and the people involved in the interaction, we predict that participants will show a tendency to mention the person of their own gender first for woman/man dyadic interactions and the person of their own race first for Black person/white person dyadic interactions. We expect syntax to modulate these effects, with stronger mention order biases manifesting when participants produce transitive trials rather than conjoined trials.

2.3. Experiment 1

2.3.1. Method

We preregistered the experiment on osf.io (https://osf.io/bfhrn). There, we specified the number of participants we would test of each demographic group, how we would code responses, the main analyses and exclusion criteria.

2.3.1.1. Participants

Using Prolific.com, we paid 240 English-speaking adults £4 to take part in this 30 minute study. 60 participants reported that they were Black women, 60 Black men, 60 white women, and 60 white men, mean reported age was 30 (range 18-70).

As participants reported their sex as either male or female and their gender identity as either male or female, we will use the language of “female” and “male” to refer to both the gender and sex of our participants, although we recognise this is not the only language that can be used to discuss gender.
2.3.1.2. Materials

Participants completed two experimental tasks: a memory task and a language task, as well as a post-study questionnaire. In both tasks, they saw photographs of the same eight figures, two Black men, two Black women, two white men and two white women (Figure 1). Each figure wore neutral clothing, was cropped against a white background, and was assigned a monosyllabic name (Beth, Kate, Ruth, Jane, Fred, Dave, Mike and Luke), counterbalancing within gender across participants.

Figure 1

An example of the memory task matrix used in Experiment 1

Fred  Ruth  Beth  Luke  Dave  Kate  Jane  Mike

If you have memorised all the portraits and names and are ready to start, press Continue.

Please note: Your memory of these faces will be tested 4 times during this experiment.

Note. This matrix displayed the eight figures who would appear individually in the memory task and in pairs in the language task.
In the memory task, participants viewed portrait photographs of the figures (Figure 1) and in the language task, participants saw photographs of these figures executing real-life interactions (e.g., Figure 2). We chose ten interactions that could all be described using a symmetrical predicate (Gleitman et al., 1996): pointing angrily (verb: *arguing*), waltzing (*dancing*), a boxing stance (*fighting*), hugging (*hugging*), pursing lips (*kissing*), holding hands (*marrying*), shaking hands (*meeting*), shouting (*shouting*), gesturing with hands (*talking*) and touching palms (*touching*).

Half the photographed interactions showed a woman and a man of the same race (gender trials, n=40), and half showed a Black person and white person of the same gender (race trials, n=40). The gender trials were evenly split between Black dyads and white dyads, and the race trials between male dyads and female dyads.

**Figure 2**

*Examples of the action “meeting”, showing a gender trial (left) and race trial (right)*
On half of the trials, participants were prompted to describe the photographs with a fragment of a symmetrical predicate in a transitive frame (e.g., \textit{is marrying...}) and on the remaining half with a conjoined fragment (e.g., \textit{are marrying each other}). The purpose of adding “each other” to the end of conjoined sentences was for creating a greater sense of similarity between the mentioned people, as described by Gleitman et al. (1996).

The post-study questionnaire measured participants’ identification with their gender and race, as well as their gender, race, age, location and their first language. We used the Multicomponent In-group Identification Survey developed by Leach et al. (2008), which addresses two dimensions; group-level self-investment and group-level self-definition, modified to separately measure gender ingroup identification and racial ingroup identification, with answers on a 7-point Likert scale (Appendix B).

2.3.1.3. Procedure

In the memory task, participants first viewed a portrait matrix depicting eight photographs of people and their names (Figure 1). We asked participants to memorise the portraits and names before a test, in which they viewed each portrait alone and matched the correct name from the 8 options. Participants repeated this memory task (using a different random order of trials each time) until they reached perfect accuracy, whereupon they began the language task. To ensure naming accuracy throughout experiment, participants repeated this same memory task after every 20 trials of the language task.

On each critical trial of the language task, participants viewed one of the photographed interactions, presented for 2000 ms but mirror-reversing every 500ms. Then, they saw either an active sentence fragment (n=40) or conjoined subject sentence
fragment (n=40) for 2000 ms and after this were prompted to complete the fragment based on the photograph they had seen, typing the resulting sentence into a text box. Trials were presented in random order. Before beginning the critical trials, participants viewed one example transitive trial and one example conjoined trial using dyads of both the same gender and race. After the language task, participants completed the in-group identification and demographic questionnaire.

Photographs were turned into GIFs using the magick package in R (Ooms, 2018). Piloting showed that this procedure highly reduced participants’ bias to mention the left-hand figure first. We counterbalanced which figure appeared on the left-hand-side first in the GIF stimuli across experimental lists, as well as which names were assigned to which figures. This experiment was programmed using JavaScript and the jsPsych library (de Leeuw, J. R., 2015), and data was collected on the Prolific.com platform.

2.3.2. Analysis

The study used a mixed design, involving a within-subject and within-item factor in which all participants received the same syntax manipulation, and a between-subject/within-item factor of participant identity, which can be conceptualised as either differences between participant groups (women compared to men, Black people compared to white people), or in terms of how the participant matched the photographed figures in terms of gender or race. (Note that all participants viewed the same photographs). We thus analysed how social cognition influenced mention order in two ways, corresponding to these conceptualisations: first, looking at characteristics of the speaker’s social context and second, characteristics of the speaker’s social identity.
2.3.2.1. Conventional social hierarchies

Analysis 1 looks at the proportion of trials on which participants mentioned the dominant group first. If behaviour was consistent with this hypothesis, on gender trials we would expect participants to show a man-first bias, consistent with androcentrism, indicated by a significant intercept. Furthermore, our primary variables tested if male participants would be more likely than female participants to mention the man first, indicated by a significant main effect of participant gender, and if a man-first bias would be more likely on transitive trials than on conjoined trials, indicated by a significant main effect of syntax. Our secondary variables additionally tested if a man-first bias would vary by whether the dyads were both Black or both white (a main effect of dyad race) and by whether participants were Black or white (a main effect of participant race), without predictions in either direction for both analyses. We also conducted an exploratory analysis testing the interactions between these fixed effects.

On race trials, we would expect participants to show a white-first bias, consistent with the societal white default, indicated by a significant intercept. Our primary variables tested if white participants were more likely to mention the white figure first than Black participants, indicated by a significant main effect of participant race, and if a white-first bias was more likely on transitive trials than on conjoined trials, indicated by a significant main effect of syntax. Secondary variables tested if a white-first bias varied by whether the dyads were both women or both men (a main effect of dyad gender) and by whether the participants were female or male (a main effect of participant gender), without predictions in either direction for both analyses. We also explored the interactions between these fixed effects.

For both the man-first and the white-first analyses, we restricted models to analysis of only two-way interactions. Three-way interactions routinely made these
models too complex for convergence, and our predictions did not extend to considerations for how participant demographic, syntax, and dyad demographic would interact.

2.3.2.2. The Like Me Bias

Analysis 2 tests our Like Me hypothesis, looking at the proportion of trials on which participants mentioned the figure most like themselves first, and the extent to which this varied by participant demographic (gender and race), dyad demographic and syntax type. We measured a Like Me bias by taking the proportion of gender trials on which participants mentioned the figure of their own gender first (a Like Me gender bias), and the proportion of race trials on which participants mentioned the figure of their own race first (a Like Me race bias). If the evidence supports this hypothesis, this would be indicated by a significant intercept in the model. Female participants should be more likely to mention the woman first and male participants should be more likely to mention the man first on gender trials, while Black people should be more likely to mention the Black person first and white people should be more likely to mention the white person first on race trials.

We also tested our prediction that syntax would have a main effect, and that mention order biases would be stronger on transitive sentences than on conjoined sentences. The purpose of this analysis was to determine if a Like Me gender/race bias is just as likely to occur across all four demographic groups, syntax conditions and dyads, or if there are differences between the groups and if these biases are more likely to occur on some sentence types or stimuli over others (comparing transitive/conjoined sentences and Black/White Dyads on gender trials, Female/Male Dyads on race trials).
This analysis also measured if participants would show a stronger Like Me gender bias when the person of their own gender was also the same race as themselves. Thus, we investigated differences between the types of dyads depicted in gender trials (Black Dyads = Black woman/Black man pairs, and White Dyads = white woman/white man pairs). If a Like Me gender bias is modulated by intersectional effects, Black women and men should show a stronger Like Me gender bias on Black Dyads than on White Dyads, while white women and men should show a stronger bias on White Dyads than on Black Dyads. We also measured if participants would show a stronger Like Me race bias when the person of their own race was also the same gender as themselves. Thus, we investigated differences between the types of dyads depicted in race trials (Female Dyads = Black woman/white woman pairs, and Male Dyads = Black man/White man pairs). If a Like Me race bias is modulated by intersectional effects, Black and white women should show a stronger Like Me race bias on Female Dyads, while Black and white men should show a stronger bias on Male Dyads.

Lastly, we investigated in-group identification scores as predictors of the Like Me biases, entering participants’ total scores on the two surveys from the Gender In-group Identification questionnaire ((Group-Level) Self-Investment and (Group-Level) Self-Definition) as predictors in the Like Me gender model, and their total scores on the race variations of these two surveys in the Like Me race model. Additionally, we ran a separate model adding participants’ scores on each item from these surveys as additional fixed effects. Items were Solidarity, Satisfaction, Centrality, Individual Self-Stereotyping and In-group Homogeneity (Leach et al., 2008).
2.3.2.3. **Statistical Models**

Our analyses used mixed effects logistic regression with mention order as the dependent variable (i.e., man first on gender trials and white first on race trials for the hierarchical analyses, Like Me for the second analyses). We took an iterative approach to model building: for gender trials, this started by looking at how our primary variables of participant gender and syntax affected mention order preferences, and then examining how this interacted with the race of the dyads, participant race, and participants’ gender ingroup identification scores. For race trials, this started by looking at how our primary variables of participant race and syntax affected word order, followed by examining how this interacted with the gender of the dyads, participant gender, and participants’ race ingroup identification scores. We contrast coded participant gender, participant race, dyad gender, dyad race and syntax and dummy coded valence for all analyses.

All analyses were performed using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). All models included by-item and by-subject random intercepts and the maximal random effects structure that permitted model convergence. If models did not converge, we removed random effects that accounted for the least variance until convergence. Only the final models are reported in the results. We describe each regression analysis using the lme4 model syntax. Discussions will only focus on significant and marginally significant results for simplicity of exposition.

2.3.2.4. **Control analyses**

We tested whether the size of any Like Me bias varied based on participant age and also whether the bias changed during the experiment’s start to its end. Finally, we tested whether participants were more accurate at remembering the names of the
demographic groups that they matched, which could be a potential explanation of any bias in mention order. We thus examined data from the first round of the memory task, using a mixed effects model to determine if accuracy was predicted by participant gender and race, and their interaction with stimuli gender and race.

2.3.2.6. Exclusions

We used stringent exclusion criteria to ensure that participants’ Prolific-reported demographic background matched their self-reported demographic background, and that they completed the experiment correctly. We excluded 5 participants whose reported race did not match with their Prolific records, and 8 participants who reported that they did not speak English as a first language.

The remaining exclusion criteria were all pre-registered. Firstly, we excluded individual trials where participants made response errors (e.g., using incorrect names, verbs, or syntax), comprising 19% of total trials. Trials where participants made a clear typing error that did not signify a figure identification or syntax use error were hand corrected and included (e.g., the participant typed “Miike” or “is huging”). Next, we excluded 31 participants for making response errors on at least 50% of transitive trials and 50% of conjoined trials.

We also excluded participants who showed a strong orientation bias, naming the left figure first on more than 90% or fewer than 10% of trials. Importantly, our mirror-reversing images appeared to strongly reduce any orientation bias, as shown in Figure 3, such that only 2 participants were excluded (mean orientation bias=54% [SD=0.13]). We suggest using this mirror-reversal method for similar picture-description tasks. Finally, we excluded trials where responses were very slow (greater than 3 standard deviations from the mean, 2% of trials).


Figure 3

*Orientation bias by participant*

![Graph showing orientation bias by participant]

*Note.* Participants were excluded if they mentioned the left-hand figure first on more than 90% of trials (n=2) or fewer than 10% of trials (n=0).

Some participants were excluded for multiple criteria. Thus, the main analysis included 196 participants (46 Black women, 45 Black men, 54 white women, 51 white men, mean age = 30.52 (SD = 9.62), range 18 – 70).

For the questionnaire analysis, we further excluded participants who routinely selected the same response on over 90% of items on either the gender in-group or racial in-group identification measures (or both). Additionally, we excluded participants who did not complete the relevant questionnaire in full (e.g., if a participant only answered 80% of the questions on the racial in-group identification questionnaire, they were excluded from the Like Me race analysis using scores from this questionnaire as a fixed
Thus, the gender in-group analysis was conducted with 158 participants (36 Black women, 32 Black men, 42 white women, 48 white men, mean age = 30.56 (SD = 9.50) and the race in-group analysis was conducted with 159 participants (36 Black women, 32 Black men, 46 white women, 45 white men, mean age = 30.4 (SD = 10.09).

One item, “It is pleasant to be my race” from the Satisfaction measure (part of group-level Self-Investment) was excluded for demonstrating an item-total correlation of under 0.5 Cronbach's $\alpha$ (0.39).

2.3.3. Results

First, we analysed data from the gender trials, testing effects of androcentrism on mention order (i.e., a man-first bias) and then our Like Me gender hypothesis. Next, we analysed data from the race trials, testing the effects of the white default on mention order (i.e., a white-first bias) and then our Like Me race hypothesis.

2.3.3.1. Man-first

Firstly, we analysed if participants showed a tendency to mention the man first on gender trials, and the extent to which mention order was influenced by participant demographic, syntax and the race of the dyads (see Figure 4 for effects on the man-first bias).
Figure 4

The mean proportion of gender trials on which the man was mentioned first

Note. Data is also distinguished by dyad race (Black Dyads = Black woman/Black man dyads, White Dyads = white woman/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the man, below it represents a bias in favour of the woman. Error bars represent standard error.

We conducted a mixed effects logistic regression of the form Man First ~ (Participant Gender + Syntax + Dyad Race + Participant Race) ^2 + (1 + Syntax | Subject) + (1 | Event). The results from this regression are displayed in Table 1.
Table 1

Results from man-first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.18</td>
<td>0.06</td>
<td>3.03</td>
<td>0.00 **</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.29</td>
<td>0.05</td>
<td>-5.55</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.06</td>
<td>0.07</td>
<td>-0.90</td>
<td>0.37</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.21</td>
<td>0.03</td>
<td>7.57</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.05</td>
<td>0.05</td>
<td>-1.01</td>
<td>0.31</td>
</tr>
<tr>
<td>Participant gender : Syntax</td>
<td>-0.14</td>
<td>0.07</td>
<td>-2.14</td>
<td>0.03 *</td>
</tr>
<tr>
<td>Participant gender : Dyad race</td>
<td>0.07</td>
<td>0.03</td>
<td>2.40</td>
<td>0.02 *</td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>-0.05</td>
<td>0.05</td>
<td>-0.99</td>
<td>0.32</td>
</tr>
<tr>
<td>Syntax : Dyad race</td>
<td>-0.04</td>
<td>0.03</td>
<td>-1.52</td>
<td>0.13</td>
</tr>
<tr>
<td>Syntax : Participant race</td>
<td>0.02</td>
<td>0.07</td>
<td>0.26</td>
<td>0.80</td>
</tr>
<tr>
<td>Dyad race : Participant race</td>
<td>0.04</td>
<td>0.03</td>
<td>1.32</td>
<td>0.19</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001.

With these stimuli, participants were overall more likely to mention the man first rather than the woman (M = 0.54 [SD = 0.20]), confirmed by a significant intercept for the regression. Critically, mention order was also affected by participant gender. Male participants mentioned the man before the woman more often than female participants did (M = 0.60 [SD = 0.19] vs M = 0.48 [SD = 0.18]).

Next, we analysed the effect of grammar. We found a significant interaction between syntax and participant gender; male participants mentioned the man first more often on transitive trials (M = 0.61 [SD = 0.20]) than on conjoined trials (M = 0.58 [SD = 0.23]), while female participants mentioned the man first less often on transitive trials (M = 0.44 [SD = 0.26]) than conjoined trials (M = 0.52 [SD = 0.24]). Syntax alone did
not have a significant effect on mention order (transitive $M = 0.53 \ [SD = 0.25]$ vs. conjoined $M = 0.55 \ [SD = 0.24]$).

Mention order was affected by the race of the dyads. Participants were significantly more likely to mention the man first when both figures were white rather than when they were Black. We also found a significant interaction between participant gender and dyad race; female participants were more likely to mention the man first when both people were white ($M = 0.54 \ [SD = 0.15]$) rather than Black ($M = 0.43 \ [SD = 0.20]$), whereas male participants were likely to mention the man first for both types of dyad (white dyads $M = 0.63 \ [SD = 0.19]$, Black dyads $M = 0.56 \ [SD = 0.19]$).

2.3.3.2. **Like Me Gender bias**

Next, we investigated if the mention order behaviour of participants was consistent with our Like Me gender bias hypothesis; Figure 5 displays the effects on this bias. To do so, we conducted a mixed effects logistic regression of the form Like Me $\sim$ Participant Gender + Syntax + Dyad Race + Participant Race $+$ (1 $+$ Syntax $+$ Dyad Race $|$ Subject). Random effects by item were removed for model convergence. The results are displayed in Table 2.
Figure 5

*The mean proportion of gender trials demonstrating a Like Me gender bias*

*Note.* Showing means of participants from all four demographic groups, on transitive and conjoined trials and across dyad race. Black Dyads = Black woman/Black man pairs, White Dyads = white woman/white man pairs. The dotted line represents chance at 50% - above the line represents a Like Me gender bias. Error bars represent standard error.
Consistent with our hypothesis of a Like Me gender bias, participants were overall more likely than chance to mention the figure of their own gender first (M = 0.56 [SD = 0.15]), as indicated by a significant intercept. We found a significant main effect of participant gender; male participants showed a stronger Like Me gender bias (M = 0.60 [SD = 0.19]) than female participants (M = 0.52 [SD = 0.18]). One-sample t-tests indicated that the Like Me gender bias was significant for male participants (t(95) = 6.25, p < .001), but not for female participants (t(99) = 1.29, p = .10).

We also found a significant main effect of syntax; participants were more likely to demonstrate a Like Me gender bias on transitive sentences (M = 0.58 [SD = 0.23]) than on conjoined sentences (M = 0.53 [SD = 0.24]).

Lastly, we found a marginally significant effect of dyad race; the Like Me gender bias was slightly stronger on trials with two Black people (M = 0.57 [SD = 0.19]) than on trials with two white people (M = 0.54 [SD = 0.19]).
Participant gender. To determine if the Like Me gender bias was present in both female and male populations, we conducted a mixed effects logistic regression on each group independently. These were of the form Like Me ~ Syntax + Dyad Race + Participant Race + (1 + Syntax + Dyad Race | Subject) + (1 | Event). The results of both models are shown in Table 3.

Table 3

Male and female results from the Like Me gender bias linear regressions

<table>
<thead>
<tr>
<th></th>
<th>Male participants</th>
<th>Female participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.48</td>
<td>0.09</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.15</td>
<td>0.06</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001. * = marginally significant.

For male participants, the intercept was significant, indicating that they were reliably showing a Like Me gender bias to mention the man first on these trials. In comparison, for female participants the intercept was non-significant, consistent with our finding that men showed a stronger Like Me gender bias than women.

We found a significant main effect of dyad race for both male and female participants. Male participants were more likely to mention the man first when dyads
were white (M = 0.63 [SD = 0.19]) than when dyads were Black (M = 0.56 [SD = 0.19]). Female participants were more likely to mention the woman first when dyads were Black (M = 0.57 [SD = 0.20]). Conversely, when dyads were white, they tended to mention the man first (M = 0.46 [SD = 0.15]). Syntax only had a significant effect on the Like Me gender bias of female participants and this was marginal.

**Gender in-group identification.** Next, we investigated how individual differences may have influenced the Like Me gender bias by entering the totals of the two surveys from the Gender In-group Identification questionnaire ((Group-Level) Self-Investment and (Group-Level) Self-Definition) as scaled fixed effects in the model, of the form Like Me ~ Participant Gender + Syntax + Dyad + Participant Race + Self-Investment + Self-Definition + (1 + Syntax + Dyad | Subject). By-item random effects were removed for model convergence. Neither Self-Investment (p = .29) nor Self-Definition (p = .49) predicted the Like Me gender bias. With the addition of these fixed effects, syntax was no longer a significant main effect (p = .20), nor was dyad a marginally significant main effect (p = .29).

Control analyses tested the effects of trial progression and participant age, but did not find strong evidence that either of these had a strong effect on the Like Me gender bias.² Next, we analysed data from the race trials.

---

² To determine if the Like Me gender bias changed over the course of the experiment, we added scaled trial number as a predictor to the main Like Me Gender model. We found a marginally significant main effect of trial number (p = .08), suggesting that the Like Me gender bias slightly increased in strength as the experiment progressed. Participant age was not a significant predictor of the Like Me gender bias (p = .71), as indicated by the addition of age as a fixed effect in the main Like Me gender model.
2.3.3.3. White-first

Firstly, we analysed if participants showed a tendency to mention the white figure first on race trials, and the extent to which mention order was influenced by participant demographic, syntax and the gender of the dyads (see Figure 6 for effects on the white-first bias).

Figure 6

*The mean proportion of race trials on which the white person was mentioned first*

![Figure 6](image)

*Note.* Means for both white and Black participants and on transitive and conjoined sentences. Data is also distinguished by dyad gender (Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the white person, below it represents a bias in favour of the Black person. Error bars represent standard error.
To do so, we conducted a mixed effects logistic regression of the form White First ~ (Participant Race + Syntax + Dyad Gender + Participant Gender)^2 + (1 + Syntax | Subject) + (1 | Event). Results from this model are shown in Table 4.

Table 4

Results from white-first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.02</td>
<td>0.06</td>
<td>-0.41</td>
<td>0.68</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.20</td>
<td>0.04</td>
<td>5.54</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.01</td>
<td>0.06</td>
<td>-0.22</td>
<td>0.83</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>-0.33</td>
<td>0.03</td>
<td>-12.47</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.89</td>
<td>0.37</td>
</tr>
<tr>
<td>Participant race : Syntax</td>
<td>0.03</td>
<td>0.06</td>
<td>0.45</td>
<td>0.65</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>0.00</td>
<td>0.03</td>
<td>0.17</td>
<td>0.87</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.38</td>
<td>0.70</td>
</tr>
<tr>
<td>Syntax : Dyad gender</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>-0.08</td>
<td>0.06</td>
<td>-1.39</td>
<td>0.17</td>
</tr>
<tr>
<td>Dyad gender : Participant gender</td>
<td>-0.03</td>
<td>0.03</td>
<td>-1.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>

***p < .001.

The intercept was not significant, showing that participants did not overall show a bias towards mentioning the white figure first more often than the Black figure with these stimuli (M = 0.50 [SD = 0.12]). However, there was a main effect of participant
race, with white participants more likely than Black participants to mention the white figure first (M = 0.53 [SD = 0.18] vs. M = 0.45 [SD = 0.17]). Thus, just as men and women differed in their likelihood to mention the more socially dominant figure first in a gendered context, so did Black and white people in a racial context.

We found a significant main effect of dyad; participants were more likely to mention the white figure first when dyads were men than when they were women (M = 0.57 [SD = 0.15] vs. M = 0.42 [SD = 0.17])

2.3.3.4. Like Me Race bias

Next, we investigated if the mention order behaviour of participants was consistent with our Like Me race bias hypothesis; Figure 7 shows effects on this bias. To do so, we used a mixed effects logistic regression of the form Like Me ~ Participant Race + Syntax + Dyad Gender + Participant Gender + (1 + Syntax + Dyad Gender | Subject) + (1 | Event). The results of this model are displayed in Table 5.

A significant intercept indicated that participants were more likely than chance to mention the figure of their own race first (M = 0.54 [SD = 0.11]), which is consistent with our Like Me race bias hypothesis. Importantly, there was no effect of syntax or participant race on the Like Me race bias. No other fixed effects were significant.
Figure 7

*The mean proportion of race trials demonstrating a Like Me race bias*

Note. Showing means of participants from all four demographic groups, on transitive and conjoined trials and across dyad gender. Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads. The dotted line represents chance at 50% - above the line represents a Like Me race bias. Error bars represent standard error.
Table 5

Results from the Like Me race bias linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.20</td>
<td>0.04</td>
<td>5.10</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.02</td>
<td>0.04</td>
<td>0.45</td>
<td>0.65</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.03</td>
<td>0.06</td>
<td>0.40</td>
<td>0.69</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>-0.02</td>
<td>0.04</td>
<td>-0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.35</td>
<td>0.73</td>
</tr>
</tbody>
</table>

***p < .001.

**Participant race.** To determine if the Like Me race bias was indeed present in both Black and white populations, we conducted a mixed effects logistic regression on each group independently. For both groups, these were of the form Like Me ~ Syntax + Dyad Gender + Participant Gender + (1 + Syntax + Dyad Gender | Subject) + (1 | Event). Results from these models are displayed in Table 6.
For both Black and white participants, the intercept was significant, indicating that both groups were reliably showing a Like Me race bias.

We also found a significant main effect of dyad gender for both groups; Black participants were significantly more likely to demonstrate a Like Me race bias when dyads were women ($M = 0.62$ [SD = 0.15]) than when dyads were men ($M = 0.48$ [SD = 0.15]). Conversely, white participants were significantly more likely to show a Like Me race bias when dyads were men ($M = 0.61$ [SD = 0.15]) than when dyads were women ($M = 0.46$ [SD = 0.17]).

**Racial in-group identification.** Next, we investigated how individual differences may have influenced the Like Me race bias. To do so, we entered participants’ scores from the in-group identification sub-scales for race, each as a scaled fixed effect, into a new model of the form Like Me $\sim$ Participant Race + Syntax + Dyad...
Gender + Participant Gender + Self-Investment + Self-Definition + (1 + Syntax + Dyad Gender | Subject). Random effects by item were removed for model convergence.

We did not find significant main effects of either racial in-group self-investment ($\beta = -0.04, SE = 0.06, z = -0.66, p = 0.51$) or self-definition ($\beta = 0.07, SE = 0.04, z = 1.60, p = 0.11$).

We ran a second model, adding participants’ scores on all five survey items as predictors to the main Like Me race model (Solidarity, Satisfaction, Centrality, Individual Self-Stereotyping, In-group Homogeneity). The results of this model are shown in Table 7.

We found a significant main effect of Individual Self-Stereotyping. Participants who scored higher on the questions “I have a lot in common with the average person of my race” and “I am similar to the average person of my race” were more likely to demonstrate a Like Me race bias than those who did not rate highly on these questions. No other items influenced the Like Me race bias.
Table 7

*Results from the Like Me race bias model with racial in-group identification items added as fixed effects*

<table>
<thead>
<tr>
<th>Estimate</th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.22</td>
<td>0.04</td>
<td>5.47</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.02</td>
<td>0.07</td>
<td>-0.33</td>
<td>0.74</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.03</td>
<td>0.07</td>
<td>0.37</td>
<td>0.71</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>-0.06</td>
<td>0.05</td>
<td>-1.31</td>
<td>0.19</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.02</td>
<td>0.04</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>Solidarity</td>
<td>-0.04</td>
<td>0.07</td>
<td>-0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.05</td>
<td>0.08</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>Centrality</td>
<td>-0.05</td>
<td>0.07</td>
<td>-0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>Individual Self-Stereotyping</td>
<td>0.12</td>
<td>0.05</td>
<td>2.53</td>
<td>0.01 *</td>
</tr>
<tr>
<td>In-group Homogeneity</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.83</td>
<td>0.40</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001.

Control analyses tested the effects of trial progression and participant age, but we did not find strong evidence that either of these had a strong effect on the Like Me race bias.³

³ To determine if the Like Me race bias changed over the course of the experiment, we added scaled trial number as a predictor to the main Like Me race model. We found a marginally significant main effect of trial number (p = .07), suggesting that this bias decreased in strength slightly as the experiment progressed. We entered participant age as a fixed effect in the main Like Me race model, but it was not significant (p = .77).
Results from this study motivated two additional exploratory analyses: valence and memory for names.

### 2.3.3.5. Valence effects

Firstly, we were able to informally categorise the events shown in this task into three valence types; intimate events (dancing, hugging, kissing, marrying), casual events (meeting, talking and touching) and negative events (arguing, fighting, shouting). We tested if valence had an effect on mention order by adding this as a dummy coded predictor in the main models.

On gender trials, we found that the tendency to mention the man first varied across trials of different valences ($\beta = 0.14$, $SE = 0.03$, $z = 4.21$, $p < .001$). Participants were more likely to mention the man first on intimate trials (M = 0.56 [SD = 0.18]) than on casual trials (M = 0.53 [SD = 0.19]) or negative trials (M = 0.51 [SD = 0.19]). Additionally, we found a significant interaction between valence and syntax ($\beta = 0.10$, $SE = 0.03$, $z = 3.15$, $p < .01$). Figure 8 indicates that participants showed a greater likelihood to mention the man first on transitive trials when they were describing intimate events, whereas they showed a similar man-first bias on conjoined trials across the three valences.

On race trials, we found a significant main effect of valence ($\beta = -0.17$, $SE = 0.03$, $z = -5.37$, $p < .001$). Participants were more likely to mention the white figure first on negative trials (M = 0.54 [SD = 0.17]) compared to casual trials (M = 0.50 [SD = 0.17]) and intimate trials (M = 0.47 [SD = 0.15]). Valence was not a significant predictor of the Like Me gender or race biases.
Figure 8

**Effects of valence on the man-first bias**

![Bar chart showing the effects of valence on the man-first bias.](Image)

*Note.* The mean proportion of gender trials on which participants demonstrated a man-first bias, on transitive and conjoined sentences and across the valence condition. The dotted line represents chance at 50% - above the line represents a bias in favour of the man, below it represents a bias in favour of the woman. Error bars represent standard error.

### 2.3.3.6. Memory effects

We explored the potential effects of memory, examining participants’ accuracy at remembering the names of the eight photographed figures by analysing mean trial success across participants during the first round of the memory task (i.e., how many
trials out of 8 they got right first time). Overall, participants were demonstrating relatively high accuracy on this task (Black women: M = 0.94 [SD = 0.14]; Black men: M = 0.94 [SD = 0.15]; white women: M = 0.93 [SD = 0.15]; white men: M = 0.95 [SD = 0.13]).

To determine if participants were better at memorising certain stimuli over others, we used a mixed effects logistics regression of the form Accuracy ~ Participant Gender * Stimulus Gender + Participant Race * Stimulus Race + (1 | Subject). Random effects by item were removed for model convergence, as were non-significant interactions.

We found a marginal effect of stimulus gender on memory task accuracy ($\beta = -0.16$, $SE = 0.09$, $z = -1.77$, $p = .08$); participants were slightly better at remembering names for men than women. Importantly, however, no interactions were significant, suggesting that there was no Like Me bias in terms of how participants remembered the stimuli, and thus that the process of memorisation was unlikely to explain any linguistic biases.

2.3.4. Discussion

This experiment sought to test if mention order for descriptions of interactions is influenced by relative social similarity. We specifically investigated gender and race, testing the degree to which mention order can be predicted by conventional social hierarchies and then by our Like Me hypothesis, that speakers will show a tendency to mention a person who is more socially similar to themselves first over a person who is less socially similar. This is the first study to address the effects of gender and race on mention order for descriptions of dyadic interactions, as well as addressing potential intersectional effects.
On gender trials, participants viewed pairs of women and men interacting and we measured if they would show a mention order bias that reflects androcentrism in their descriptions of these events, tending to mention the man first and the woman second. We then investigated if participants showed a Like Me gender bias, with women mentioning the woman first and the man second, and men mentioning the man first and the woman second. On race trials, participants viewed pairs of Black and white people interacting and we measured if they would show a mention order bias that reflects a white default in society, tending to mention the white person first and the Black person second. We then investigated if participants showed a Like Me race bias, with Black participants mentioning the Black person first and the white person second, and white participants mentioning the white person first and the Black person second.

For our gender analyses, we found that participants did generally mention the man before the woman. However, male participants were significantly more likely to do so than female participants. This effect of participant gender was reflected in the finding that the man-first trend was moderated by a Like Me gender bias, as participants tended to mention the person of their own gender first (male participants mentioning the man first and female participants mentioning the woman first). Men demonstrated a stronger Like Me gender bias than women, who in fact did not show a significant Like Me gender bias.

For our race analyses, participants were not generally more likely to mention the white person before the Black person, and so the conventional social hierarchical explanation for mention order did not hold in this instance. A significant main effect of participant race indicated that white participants were much more likely to mention the white person first than Black participants. Our data instead supported our Like Me race hypothesis, as participants showed a tendency towards mentioning the person of their
own race first. The Like Me race bias did not differ between Black and white participants.

The results from this experiment therefore support our predictions that people will demonstrate a Like Me bias in their language production, tending to mention the figure of their own gender first, and the figure of their own race first when describing dyadic interactions. These results highlight a crucial role for the speaker’s identity in production biases and contribute to a growing body of research that has identified closeness to the speakers as an influential factor of mention order. However, this overall Like Me bias was not uniform: further inspection of descriptive statistics and results from one-sample t-tests revealed that all demographic groups demonstrated a Like Me gender bias apart from Black women (M = 0.49 [0.18], t(45) = -0.28, p = .61; vs. Black men M = 0.59 [0.19], t(44) = 4.35, p < .001; white women M = 0.54 [0.19], t(53) = 1.83, p < .05; and white men M = 0.60 [0.20], t(50) = 4.46, p < .001). Furthermore, a significant overall man-first bias suggests that there was both a Like Me Gender effect and an influence of social hierarchy on mention order for gender trials, with women as well as men tending to mention men before women.

We had expected Like Me biases to appear stronger on transitive trials compared to conjoined trials, as the former involves differences in thematic role and of grammatical function. However, syntax had an inconsistent effect across gender and race trials. On gender trials, participants did show a stronger Like Me gender bias on transitive trials, however subsequent analyses suggested that this was mostly driven by female participants. On race trials, however, syntax did not reliably influence the Like Me race bias. Syntax was not a significant predictor of either a man-first or white-first bias.
We did not find evidence for intersectional effects on the Like Me Gender bias and Like Me Race bias, as indicated by patterns in Figures 5 and 7 respectively. Our findings did not suggest that Black women and men showed a stronger Like Me Gender bias on trials depicting two Black people rather than two white people, or that white women and men showed a stronger Like Me Gender bias on trials depicting two white people rather than two Black people, nor did we find evidence that Black and white women showed a stronger Like Me Race bias on trials depicting two women rather than two men, or that Black and white men showed a stronger Like Me Race bias on trials depicting two men rather than two women.

Results revealed unexpected effects of dyad on mention order. On gender trials, participants were more likely to show a man-first bias when the dyads depicted a white woman and a white man together, rather than a Black woman and Black man. On race trials, participants were more likely to show a white-first bias when the dyads depicted a Black man and a white man, rather than a Black woman and a white woman. These findings may indicate that there were possible influences of visual differences (outside of gender or race) that may have contributed to these effects. We must therefore consider how variation between the figures in our dyads may have contributed to these effects of dyad identity. Specifically, the male figures presented were generally taller than the female figures, and while some figures were dressed in bright or white colours, others were in black clothing (Figure 1). As photographs were of real-life interactions involving the actors physically touching, we were not able to standardise the visual salience of the depicted figures – for example, by resizing figures to be the same height. Furthermore, facial expressions were not standardised between the depicted pairs, with some figures smiling more than their partner, for example.
Thus, in the next study we aimed to replicate the finding of a Like Me gender and race bias using a new set of photographic stimuli that were controlled for irrelevant visual characteristics. This would allow us to argue more confidently in favour of a Like Me bias, to further examine the potential influence of syntax, and to interrogate the effects of dyad using a more standardised set of stimuli.

Exploratory results also revealed effects of valence (intimate, casual and negative events). Mention order biases have already been associated with the negatively valanced topic of war and opposition, with speakers more likely to mention their in-group ahead of an out-group when naming conflicts (Oeberst & Matschke, 2017). Conflict offers an interesting case study of mention order because there can be clear “sides”. However, social language is not only conflict-based, but can include anything from relatively mundane interactions to highly valanced events. The emerging effects of valence in this experiment suggest that this may be a valuable factor to explore more formally in our investigations of mention order for describing interactions, and we will do so in following experiments and chapters.

Lastly, we will continue to assess participants’ accuracy on the memory task, to help rule out the explanation that the Like Me bias could be attributed to participants showing a better ability for remembering the names of the people of their own gender or of their own race.

2.4. Experiment 2

Experiment 2 attempted to replicate the main findings of Experiment 1, but with more standardised stimuli. We preregistered the experiment on osf.io (https://osf.io/t928m).
In these new stimuli, we specifically controlled for a range of potential varying factors. This included clothing, with all actors wearing dark trousers and different coloured pastel shirts; height, with actors all a similar height and not touching during interactions, making resizing possible; and facial expression, which was similar across dyads so that if one figure was smiling, frowning, or neutral, so was the other figure.

Figure 9 depicts the new portrait matrix.

**Figure 9**

*An example of the memory task matrix from Experiment 2*

![Figure 9](image_url)

*Note.* This matrix displayed the eight figures who would appear individually in the memory task and in pairs in the language task.
In addition, Experiment 2 did not measure in-group identification, but instead considered how the nature of the events in Experiment 1 may have elicited different responses in participants due to attitudes towards particular social groups. Specifically, we considered if the speaker’s attitudes toward same-gender relationships and interracial relationships may have influenced responses to intimate events between people of the same gender and between people of different races. We chose to investigate how these attitudes may influence the Like Me race bias, predicting that divergences from predictable patterns of behaviour (i.e., mentioning the figure of their own race first) would most likely occur on trials that contradicted the speaker’s expectations and accepted world view (e.g., two men shown “marrying”, described by a participant with negative attitudes towards homosexuality).

Thus, we included a post-study questionnaire to address attitudes towards same-gender relationships and interracial relationships. Attitudes towards same-gender relationships were measured using an item from the European Social Survey (2002–2010) and items from Watt and Elliot (2019). Attitudes towards interracial relationships were measured using items from Field, Kimuna and Straus (2013).

2.4.1. Method

Methods were as in Experiment 1, unless specified otherwise.

2.4.1.1. Participants

240 new participants were recruited from Prolific.com, using the same principles as Experiment 1 (60 white men, 60 Black men, 60 white women, 60 Black women).
2.4.1.2. **Materials**

We recruited a new set of individuals for our photographs. Again, they were photographed as standing portraits, as well as taking part in actions. However, rather than having them directly interact with each other, we only used events where the individuals did not strictly touch, so that their photographs could be separately edited. This meant that we replaced two of the symmetrical predicates, *hugging* and *meeting*, with *flirting* and *playing cards*.

Attitudes towards same-gender relationships were measured using three items: “Gay men and lesbians should be free to live their own life as they wish”, ESS 2002–2010; “What are your opinions about sexual relations between two adult men?”; “What are your opinions about sexual relations between two adult men?” (Watt & Elliot, 2019). Attitudes towards interracial relationships were also measured using three items: “I think it is good for Black and White people to be friends”; “I think it is good for Black and White people to date”; “I think it is good for Black and White people to marry” (Field et al., 2013).

Analyses were conducted using the same approach as Experiment 1. As before, we analysed if mention order was affected by fixed social hierarchies (man-first on gender trials, white-first on race trials) and if they tended to show a Like Me bias to mention the person of their own gender first, and the person of their own race first. We analysed gender trials and then race trials. This time, we also included valence as a fixed effect. Trials were categorised as intimate events (dancing, flirting, kissing, marrying), casual events (playing, talking and touching) and negative events (arguing, fighting, shouting).
2.4.1.3. **Exclusions**

We used the same exclusion criteria as in Experiment 1 for individual participants and individual trials. We excluded 5 participants whose reported race did not match across the two records, 2 participants whose reported gender did not match and 7 participants with a mismatch in whether they reported being L1 English speakers.

The remaining exclusion criteria were all pre-registered. 18% of total trials were excluded for response errors and 26 participants were excluded for making response errors on 50% more of transitive trials, or 50% or more of conjoined trials. 5 participants were excluded for showing a strong orientation bias, naming the left figure first on more than 90% of trials. Again, we found that using mirror-reversing images avoided significant orientation biases across participants (mean of 53% [SD=0.15]). 1% of trials were excluded for reactions times greater than 3 standard deviations from the participant mean. After exclusions, the main analyses included 195 participants (51 Black women, 42 Black men, 54 white women, 48 white men; mean age = 33.16 (SD = 11.34), range 18 – 66).

For the questionnaire analysis, we further excluded trials with missing responses and excluded participants who routinely selected the same response on over 90% of items on either the attitudes towards same-gender relationships or the attitudes toward interracial relationships measures, or both. Additionally, we excluded participants from any analyses utilising these measures as predictors of the Like Me race bias, if those participants did not complete the relevant questionnaire in full.

For the attitudes toward same-gender relationships analysis, one item was removed from the dataset during a reliability check (“Gay men and lesbians should be free to live their own life as they wish” – ESS, 2002–2010), for demonstrating a very low item-total correlation (Cronbach's α = 0.49). We excluded participants for not
completing all the items in the questionnaire, leaving 59 participants remaining in this analysis (17 black men, 20 black women, 16 white men, 6 white women; mean age = 36.58; range = 18 - 64 (SD = 12.28).

For the attitudes towards interracial relationships analysis, one item was removed from the dataset (“I think it is good for Black and White people to be friends” - Field, Kimuna & Straus, 2013), for demonstrating a very low item-total correlation (Cronbach's α = 0.49). After exclusions, this analysis included 82 participants (26 black men, 27 black women, 19 white men, 10 white women; mean age = 36.32; range = 18 - 66 (SD = 12.82).

2.4.2. Results

As in Experiment 1, first we analysed data from the gender trials, testing effects of androcentrism on mention order (i.e., a man-first bias) and then the Like Me gender hypothesis. Next, we analysed data from the race trials, testing the effects of a white default in society on mention order (i.e., a white-first bias) and then the Like Me race hypothesis.

2.4.2.1. Man-first

We analysed if participants showed a tendency to measure the man first and the woman second, and the factors that may have affected mention order (see Figure 10 for effects on the man-first bias). To do so, we used a mixed effects logistic regression of the form Man First ~ (Participant Gender + Syntax + Participant Race + Dyad Race + Valence) ^2 + (1 + Syntax | Subject) + (1 | Item). Table 8 shows the results.
Figure 10

The mean proportion of gender trials on which the man was mentioned first

Note. Means for both male and female participants and on transitive and conjoined sentences. Data is also distinguished by dyad race (Black Dyads = Black woman/Black man dyads, White Dyads = white woman/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the man, below it represents a bias in favour of the woman. Error bars represent standard error.
Replicating Experiment 1, a significant intercept indicated that participants were overall more likely to mention the man first rather than the woman ($M = 0.52$ [SD = 0.17]). Critically, we again found that participant gender was a significant predictor of
mention order. Male participants were more likely to mention the man before the woman than female participants (M = 0.57 [SD = 0.16] vs. M = 0.49 [SD = 0.16]).

Thus, the key findings from Experiment 1 were replicated. Additionally, we found a main effect of valence, with participants showing a greater tendency to mention the man first on intimate trials (M = 0.55 [SD = 0.17]) compared to casual trials (M = 0.52 [SD = 0.19]) or negative trials (M = 0.49 [SD = 0.18]).

For other findings, however, Experiment 2 provided quite distinct results. First, while there was a null effect of syntax in Experiment 2 (as in Experiment 1), there was no interaction between syntax and participant gender. Second, using these new stimuli, we found a significant main effect of dyad race, but this was in the opposite direction to Experiment 1. This time, participant race had a significant effect; Black participants were more likely to mention the man first than white participants (M = 0.55 [SD = 0.14] vs. M = 0.50 [SD = 0.12]).

We also found a significant interaction between dyad race and syntax. Patterns depicted in Figure 11 indicate that, while constructing conjoined sentences, participants showed a tendency to mention the woman first when both figures were white. This is compared to transitive trials, on which participants often mentioned the man first regardless of dyad race. A marginal interaction emerged between participant race and valence.
2.4.2.2. **Like Me Gender bias**

Next, we investigated if participants were demonstrating a Like Me bias to mention the person of their own gender first (see Figure 12 for effects on the Like Me gender bias). To do so, we conducted a mixed effects logistic regression of the form

\[
\text{Like Me} \sim \text{Participant Gender} + \text{Syntax} + \text{Participant Race} + \text{Dyad Race} + \text{Valence} + (1
\]
+ Syntax + Dyad Race + Valence | Subject), with results displayed in Table 9. Random effects by item were removed to avoid a singular fit.

**Figure 12**

*The mean proportion of gender trials demonstrating a Like Me gender bias*

![Figure 12](image)

*Note.* Showing means of participants from all four demographic groups, on transitive and conjoined trials and across dyad race. Black Dyads = Black woman/Black man dyads, White Dyads = white woman/white man dyads. The dotted line represents chance at 50% - above the line represents a bias in favour of the Like Me Gender bias. Error bars represent standard error.
Table 9

Results from the Like Me gender bias linear regressions

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.19</td>
<td>0.04</td>
<td>4.47</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.11</td>
<td>0.05</td>
<td>-2.41</td>
<td>0.02 *</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.08</td>
<td>0.06</td>
<td>-1.33</td>
<td>0.18</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.04</td>
<td>0.03</td>
<td>1.33</td>
<td>0.19</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>Valence</td>
<td>0.01</td>
<td>0.03</td>
<td>0.29</td>
<td>0.77</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001.

A significant intercept indicated that, overall, participants were more likely to mention the figure of their own gender first rather than the figure of a different gender (M = 0.54 [SD = 0.13]). A significant main effect of participant gender indicated that male participants showed a stronger Like Me gender bias than female participants (M = 0.57 [SD = 0.16] vs M = 0.51 [SD = 0.16]). One-sample t-tests indicated that the Like Me gender bias was significant for male participants (t(89) = 5.02, p < .001), but not for female participants (t(104) = 1.03, p = .15). We did not find significant main effects of syntax or dyad race on the Like Me gender bias.

We replicated the finding that Black women were the least likely demographic to show a Like Me gender bias; a result which accounts for the marginal female Like Me Gender bias above. One-sample t-tests showed that Black women did not demonstrate a significant Like Me gender bias, while the other three groups did (Black women M = 0.49 [SD = 0.17], t(50) = -0.76, p = .77; vs. Black men M = 0.60 [SD =
and .01); white men M = 0.54 [SD = 0.15], t(47) = 2.23, p < .01).

**Participant gender.** We conducted a mixed effects logistic regression on male and female groups independently. For the male participants, this was of the form Like Me ~ (Syntax + Participant Race + Dyad Race + Valence) ^2 + (1 + Syntax + Dyad Race + Valence | Subject) + (1 | Item). For the female participants, this was of the form Like Me ~ (Syntax + Participant Race + Dyad Race + Valence) ^2 + (1 + Syntax + Dyad Race | Subject) + (1 | Item). Results are displayed in Table 10.

### Table 10

*Male and female results from the Like Me gender bias linear regressions*

<table>
<thead>
<tr>
<th></th>
<th>Male participants</th>
<th>Female participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\beta)</td>
<td>SE</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.31</td>
<td>0.07</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>Dyad race</td>
<td>-0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.15</td>
<td>0.06</td>
</tr>
<tr>
<td>Valence</td>
<td>0.16</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001.

From these results, we only found a significant intercept for male participants. Participant race was a significant predictor in both male and female populations (Black
men showed a stronger Like Me gender bias than white men, while white women showed a stronger Like Me gender bias than Black women). Valence was also a significant predictor for both male and female participants. We only found a significant main effect of dyad race in the female population, who were more likely to show a Like Me gender bias when both figures were white.

Control analyses tested the effects of trial progression and participant age, but did not find strong evidence that either of these had a strong effect on the Like Me gender bias⁴. Next, we analysed data from the race trials.

2.4.2.3. **White-first**

We analysed if participants tended to mention the white figure first and the Black figure second, to test the effect of a fixed social hierarchy (see Figure 13 for effects on the white-first bias). To do so, we conducted a mixed effects logistic regression of the form White First ~ (Participant Race + Syntax + Participant Gender + Dyad Gender + Valence) ^2 + (1 + Syntax + Dyad Gender | Subject). Random effects by item were removed for model convergence. The results of this model are shown in Table 11.

---

⁴ We did not find significant main effects of trial number (\(p = .57\)) or participant age (\(p = .74\)) on the Like Me gender bias.
Figure 13

The mean proportion of race trials on which the white person was mentioned first

Note. Means for both white and Black participants and on transitive and conjoined sentences. Data is also distinguished by dyad gender (Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the white person, below it represents a bias in favour of the Black person. Error bars represent standard error.
Table 11

Results from white-first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.02</td>
<td>0.04</td>
<td>0.39</td>
<td>0.70</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.24</td>
<td>0.04</td>
<td>5.77</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.03</td>
<td>0.06</td>
<td>0.54</td>
<td>0.59</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.05</td>
<td>0.03</td>
<td>1.52</td>
<td>0.13</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.02</td>
<td>0.04</td>
<td>-0.41</td>
<td>0.68</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.09</td>
<td>0.03</td>
<td>-2.76</td>
<td>0.01 **</td>
</tr>
<tr>
<td>Participant race : Syntax</td>
<td>-0.04</td>
<td>0.06</td>
<td>-0.76</td>
<td>0.45</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.22</td>
<td>0.83</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>0.05</td>
<td>0.03</td>
<td>1.66</td>
<td>0.10 *</td>
</tr>
<tr>
<td>Participant race : Valence</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>Syntax : Dyad gender</td>
<td>0.02</td>
<td>0.03</td>
<td>0.75</td>
<td>0.46</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>0.06</td>
<td>0.06</td>
<td>0.96</td>
<td>0.33</td>
</tr>
<tr>
<td>Syntax : Valence</td>
<td>0.02</td>
<td>0.03</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Dyad gender : Participant gender</td>
<td>0.01</td>
<td>0.03</td>
<td>0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>Dyad gender : Valence</td>
<td>-0.05</td>
<td>0.03</td>
<td>-1.59</td>
<td>0.11</td>
</tr>
<tr>
<td>Participant gender : Valence</td>
<td>-0.07</td>
<td>0.03</td>
<td>-2.14</td>
<td>0.03 *</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001. * = marginally significant.
With these stimuli, a non-significant intercept indicated that participants were not overall more likely to mention the white person over the Black person (M = 0.50 [SD = 0.13]). However, participant race was a significant predictor of mention order; white participants were more likely than chance to mention the white figure first than Black participants (M = 0.55 [SD = 0.15] vs M = 0.45 [SD = 0.16]). We also found a significant main effect of valence; participants were more likely to mention the white person first on negative trials (M = 0.53 [SD = 0.24]) than on intimate trials (M = 0.49 [SD = 0.21]) or casual trials (M = 0.49 [SD = 0.25]).

We found a marginally significant interaction between participant race and dyad gender, however dyad gender was not a significant predictor on its own. Syntax was also non-significant. Lastly, we found a significant interaction between participant gender and valence (p < .05); Figure 14 shows that female participants were least likely to mention the man first on intimate trials compared to casual and negative trials, and instead tended to mention the woman first on these trials.
The interaction between participant gender and valence for the white-first bias

**Notes.** The mean proportion of race trials on which male and female participants demonstrated a white-first bias, across the valence condition. The dotted line represents chance at 50% - above the line represents a bias in favour of the white person, below it represents a bias in favour of the Black person. Error bars represent standard error.

**2.4.2.4. Like Me Race bias**

Next, we reanalysed these data in terms of whether participants were more likely to mention the person most like them first (see Figure 15 for effects on the Like Me race bias). To do so, we used a mixed effects logistic regression of the form Like Me ~ Participant Race + Syntax + Participant Gender + Dyad Gender + Valence + (1 + Syntax + Dyad Gender | Subject). Random effects by item were removed for model convergence. The results of this model are displayed in Table 12.
Figure 15

The mean proportion of race trials demonstrating a Like Me race bias

Note. Showing means of participants from all four demographic groups, on transitive and conjoined trials and across dyad gender. Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads. The dotted line represents chance at 50% - above the line represents a Like Me race bias. Error bars represent standard error.
### Table 12

Results from the Like Me race bias linear regression

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>( SE )</th>
<th>( z )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.23</td>
<td>0.04</td>
<td>5.76</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.01</td>
<td>0.04</td>
<td>0.25</td>
<td>0.81</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.04</td>
<td>0.06</td>
<td>-0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.06</td>
<td>0.03</td>
<td>1.70</td>
<td>0.09 *</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.25</td>
<td>0.80</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.58</td>
<td>0.56</td>
</tr>
</tbody>
</table>

***\( p < .001 \). * = marginally significant.

A significant intercept indicated that participants were more likely than chance to mention the figure of their own race first \((M = 0.55 \ [SD = 0.12])\), replicating a main finding from Experiment 1. We also found a marginally significant main effect of dyad gender.

**Participant race.** To determine if the Like Me race bias was present in both Black and white populations, we conducted mixed effects logistic regressions on both groups independently. For Black participants, this was of the form Like Me ~ Syntax + Participant Gender + Dyad Gender + Valence + (1 + Syntax + Dyad Gender | Subject). Random effects by item were removed for model convergence. For white participants, this was of the form Like Me ~ Syntax + Participant Gender + Dyad Gender + Valence
+ (1 + Syntax + Dyad Gender | Subject) + (1 | Event). Results from these models are displayed in Table 13.

Significant intercepts were found for both Black and white participants. Dyad gender was only a significant predictor of the Like Me race bias with white participants, who showed a stronger Like Me race bias on trials depicting two women (see Figure 14). We found a significant main effect of valence for white participants, who were numerically more likely to mention the white figure first on negative trials (M = 0.59 [SD = 0.24]) than on intimate trials (M = 0.53 [SD = 0.20]) or on casual trials (M = 0.54 [SD = 0.23]). The effect of valence was only marginally significant for Black participants.

Table 13

Results from the Like Me race bias linear regressions for Black and white participants

<table>
<thead>
<tr>
<th></th>
<th>Black participants</th>
<th>White participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Valence</td>
<td>0.08</td>
<td>0.05</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001. * = marginally significant.
Control analyses tested the effects of trial progression and participant age, and also affects of attitudes towards interracial and same-gender relationships on the Like Me race bias. Neither analyses showed significant effects, thus, we cannot say that participants’ mention order biases on race trials differed because of social attitudes.5

2.4.2.5. Memory effects

All four groups demonstrated relatively high accuracy at remembering the names of the eight figures (Black women: $M = 0.90$ [SD = 0.17]; Black men: $M = 0.93$ [SD = 0.15]; white women: $M = 0.93$ [SD = 0.16]; white men: $M = 0.93$ [SD = 0.14]). To analyse the interactions between participant demographic and stimulus demographic, we used a mixed effect logistics regression of the form Accuracy ~ Participant Gender + Stimulus Gender + Participant Race * Stimulus Race + (1 | Subject). All other interactions were non-significant, and so we removed these from the model for convergence along with random effects by item. Overall, participants were significantly better at remembering names for men than women ($\beta = -0.16$, $SE = 0.08$, $z = -1.99$, $p < .05$).

2.4.3. Discussion

Experiment 2 provides further evidence that people demonstrate a Like Me bias that favours the figure of their own gender and of their own race. Overall, participants

---

5 We did not find significant main effects of trial number ($p = .32$) or participant age ($p = .12$) on the Like Me race bias. We also analysed how attitudes towards two different social factors, interracial and same-gender relationships, may have influenced mention order on race trials, using scores from the post-study questionnaire. We entered scaled scores for each measure to the main model, analysing each measure separately. These models were of the form Like Me ~ Participant Race + Syntax + Participant Gender + Dyad + Valence + Interracial Relationships/Same-Gender Relationships Attitude + (1 + Syntax + Dyad | Subject). Random effects by item were removed for model convergence. We did not find significant main effects of either attitudes towards interracial relationships ($p = .56$) or same-gender relationships ($p = .70$).
showed a bias to mention the man first and the woman second but we found that male participants were driving this effect, as female participants did not show a general tendency to mention the man first. Therefore, these results do not indicate that mention order choices were driven by a man-first bias. Instead, we again found a significant Like Me gender bias and, interestingly, replicated the finding that all individual demographic groups demonstrated this bias apart from Black women. Men also showed a stronger Like Me gender bias than women.

Participants were not overall more likely to mention the white person first and the Black person second. Instead, we again found a robust Like Me race bias in all four demographic group, suggesting effects of similarity influenced mention order over effects of fixed social hierarchy.

Contrary to Experiment 1, syntax did not influence mention order. Syntax did not modulate whether participants would show a man-first or white-first bias, and participants were just as likely to mention the figure of their own gender or race first on transitive trials as they were on conjoined trials. These results suggest that differences in grammatical function are not required for a Like Me bias.

Again, we found no evidence to suggest that intersectional effects modulated either Like Me bias. As depicted in Figure 12, participants did not appear more likely to show a Like Me gender bias when the dyads were also the same race as themselves versus a different race. As depicted in Figure 15, participants also did not appear to be more likely to show a Like Me race bias when dyads were also the same gender as themselves versus a different gender.

We found a main effect of valence on both the man-first and white-first biases. Participants were more likely to mention the man first on intimate trials than on casual
trials or negative trials. They were also more likely to mention the white person first on negative trials than on intimate trials or casual trials. These results replicate the valence effects found in Experiment 1.

2.5. **Experiment 3**

An explanation for the Like Me bias could be that speakers are more likely to mention the person who is most similar to themselves first because they perceive people who are like them as more agentive than people who are less like them. While limited, there is some evidence that gender and race affect perceptions of agency. For instance, Livingston et al. (2012) discovered that (non-Black) people would ascribe a similar leadership status, a possible property of agency, to dominant Black women as to dominant white men. They found that both groups avoided a degree of backlash that dominant white women and Black men received on evaluations of leadership performance and behaviour. This finding implies that perceptions of agency are likely to be more complex than a straightforward social hierarchical explanation might suggest (e.g., the speculation that men would be perceived as more agentive than women, and white people would be perceived as more agentive than Black people). However, it is currently unclear if people show an effect akin to a ‘Like Me’ bias regarding relative perceptions of agency.

We tested this theory in a third experiment, in which we showed participants the same images used in Experiment 2 and asked them to rate who they perceived to be the agent of the event. Causation and volition are two properties of prototypical agency according to Dowty (1979, 1991). Both of these properties have been tested by Kako (2006), who found that people are more likely to rate the subject of a sentence as a causative and volitional agent than the object of a sentence. Thus, we used these two
properties to measure perceptions of agency, asking participants to rate the extent to which they perceived one figure as causing the event to occur, or wanting the event to occur.

Firstly, we considered how societal biases might influence perceptions of agency, exploring the extent to which living in a patriarchal and androcentric society encourages people to perceive men as more agentive than women, and the extent to which living in a white-majority society where whiteness is considered the default encourages people to perceive white people as more agentive than Black people, on sight alone. Secondly, we approached this from a similarity perspective reflective of the Like Me bias, exploring if participants will perceive figures who are the same gender or the same race as themselves to be the more likely agent than figures of a different gender or race.

2.5.1. Method

We preregistered the experiment on osf.io (https://osf.io/3b25f).

2.5.1.1 Participants

Using Prolific.com, we paid 128 adults £1.88 to take part. 32 participants reported that they were Black women, 32 reported that they were Black men, 32 reported that they were white women and 32 reported they were white men. Mean reported age was 32.60 [SD = 11.25], with a range of 18 – 75 years. All participants reported that their first language was English, their location as the UK, that they had no language-related disorders, and all had a Prolific approval rating of between 95% - 100%. Participants had not taken part in Experiments 1 or 2.
We conducted an effect size analysis with data from Experiments 1 and 2 to help inform us about our sample size decision. We used the mean proportion of trials on which the male figure was mentioned first on different-gender trials, comparing means for male participants with female participants, and the mean proportion of trial on which the white figure was mentioned first on different-race trials, comparing means for white participants with Black participants.

With 64 participants and an alpha of less than .05, power is 0.92 if the effect size is as in Experiments 1 and 2 on different-gender trials (d = 0.60), and power is 1 if the effect size is as in these Experiments on different-race trials (d = 0.78).

2.5.1.2. Materials

Images were presented as GIFs, depicting interactions between two figures. These events, which were the same ten symmetrical events as in Experiment 2, either showed dyads of different genders (one man, one woman) or different races (one Black person, one white person). After each GIF was presented, participants then viewed the same image above a 0 – 100 sliding scale (with numbers hidden), with a written description of the event from the perspective of one figure at point 0, and the same description but from the perspective of the other figure at point 100. These descriptions substituted names for the portraits of the figures from the event (e.g., “[portrait of Person 1] started flirting with [portrait of Person 2] first” at point (0) and “[portrait of Person 2] started flirting with [portrait of Person 1] first” at point (100); see Figure 16.
Critical trials consisted of 40 images of dyads interacting and two rating questions (one causation, one volition) per event. Statements on the rating scales measured two specific aspects of agency: causation (causing another object to experience an event or change in state) and volition (choosing to be involved in the event or state change).
Half of the critical trials were gender trials: 20 images depicted different-gender pairs, with 10 Black woman/Black man pairs and 10 white woman/white man pairs. The other half of critical trials were race trials: 20 images depicted different-race pairs, with 10 Black woman/white woman pairs and 10 Black man/white man pairs.

We used a post-study questionnaire to collect information about participants, including their age, the first language they learnt to speak, the language they currently used the most, the country they currently lived in, their nationality, their race and their gender (to check that their race and gender given in the questionnaire matched the race and gender they reported on their Prolific account), in addition to the question “Have you lived in the UK for the last 5 years or more?” with a yes/no response.

Participants’ attention during the rating task was checked using 8 catch-events, each with two ratings. On catch trials, ratings were correct or incorrect, rather than the subjective ratings of critical trials, depending on if they skewed in one direction or the other. For example, if the original phrase was "[portrait of Person 1] is chasing [portrait of Person 2]" and the first rating question showed “[portrait of Person 2] is running away from [portrait of Person 1]” at point 0 and “[portrait of Person 1] is running away from [portrait of Person 2]” at point 100, we accepted ratings from 0 – 49 as correct.

We counterbalanced across trials whether the man appeared on the left or the right first in our GIFs on different-gender trials (50% of trials were man-on-left-first, the other 50% were man-on-right-first), and whether the white person appeared on the left or the right first on different-race trials (50% of trials were white-on-left-first, the other 50% were white-on-right-first).

We also counterbalanced which figure appeared at points 0 or 100 on the rating scale as the agentive figure. Lastly, we randomised the order in which the rating
questions were presented, i.e., whether participants were first asked to rate causation or volition.

2.5.1.3. Procedure

Participants viewed images of two people interacting (e.g., person 1 and person 2 flirting with each other) and then viewed two statements describing the event in slightly different ways, from the perspective of each figure (Figure 16). They then provided ratings based on how closely they agreed with either statement. On causation trials, participants rated the degree to which they perceived one figure or the other to have caused the interaction, and on volition trials they rated the degree to which they perceived one figure or the other to have wanted the interaction to occur.

To check attention, participants answered catch questions which appeared randomly between critical trials, and after completing the language task, they completed a post-study demographics questionnaire. This experiment was programmed using JavaScript and the jsPsych library (de Leeuw, J. R., 2015).

2.5.1.4. Analysis

The design of this study is mixed, involving a within-subjects and -items design in which all participants viewed the same trials, and a between-subject/within-items factor of identity (whether the participant matched the photographed figures in gender or race). We measured perceptions of agency using participants’ responses on the rating scale. Analysis of agency perceptions was achieved by obtaining the mean scores when participants rated the dominant figure (the man on gender trials and the white figure on race trials) as being more agentive than the non-dominant figure, and whether this varied by our primary variables of participant gender (female/male), participant race
(Black/white) and dyad condition (Black Dyads/White Dyads on gender trials, Female Dyads/Male Dyads on race trials). To determine if gender influences perceptions of agency, we took a subset of the data using only the gender trials. To determine if race influences perceptions of agency, we took a subset of only race trials.

We measured agency perceptions by taking combined ratings of causation and volition, and then ratings of each agency property separately. We reverse-coded scores when the dominant figure appeared to the left of the scale so that initial scores of 0 – 49 (inclusive) were recorded as 51 – 100, favouring the dominant group, and when the non-dominant figure (the woman on gender trials, the Black person on race trials) appeared to the right of the scale, initial scores of 51 – 100 were recoded as 0 – 49. This scale was further recoded during analysis, from 0 – 100 to a scale of -50 to 50, with 0 representing the mid-point and thus chance.

Analyses investigated effects of conventional social hierarchy to determine if participants were showing a bias to perceive the man as more agentive than the woman on gender trials, and the white person as more agentive than the Black person on race trials, which would be indicated by a significant intercept. We also investigated a possible Like Me effect to determine if male participants were more likely than female participants to rate the man as more agentive than the woman on gender trials (as indicated by a significant main effect of participant gender), and if white participants were more likely than Black participants to rate the white person as more agentive than the Black person on race trials (as indicated by a significant main effect of participant race). If a Like Me bias exists on this processing task, this might offer a possible explanation for the mention order biases found in the previous experiments.

Analysis of dyad demographic explored whether participants were more likely to rate the man as more agentive than the woman on gender trials if these dyads were
Black or white, and if participants were more likely to rate the white person as more agentive than the Black person on race trials if these dyads were male or female, without strong predictions in either direction for each analysis. For both gender trials and race trials, we also examined the interaction between participant demographic and dyad demographic. For example, if white male participants would be more likely to rate the man as more agentive than the woman if the dyads were white, than if they were Black, and if Black female participants would be more likely to rate the woman as more agentive than the man if the dyads were Black rather than white (a three-way interaction between participant gender, participant race and dyad race).

All analyses were performed using linear mixed effects modelling in R (R Core Team, 2019) via the lme4 package (Bates et al., 2015). All models included by-item and by-subject random intercepts and the maximal random effects structure that permitted model convergence. If the models had a singular fit or did not converge, we removed random effects that accounted for the least variance until the models were appropriate. Only the final models are reported in the results. We describe each regression analysis using the lmeTest model syntax (Kuznetsova et al., 2017).

2.5.1.5. Exclusions

We used the same exclusion criteria for participants as in Experiment 2, in addition to excluding participants who did not report living in the UK for the last five years (7 participants), as per our pre-registration. We excluded seven participants for not being an English L1 speaker, eight participants for mis-reporting their race, and two participants for mis-reporting their gender (with some cross-over between these criteria). We calculated the overall mean trial response time and standard deviation for each participant and then excluded trials for each participant with a response time.
greater than 3 standard deviations from this mean. Analyses also excluded participants who failed more than 2 of the 16 catch questions (resulting in 64 participants excluded), and who routinely selected the same response on over 80% of trials (no participants had to be excluded for this criteria).

In our pre-registration, we stated that we would be recruiting 64 participants (16 participants of each demographic group). However, we found that almost half of participants were failing more than two catch questions and so we recruited a further 64 participants on Prolific, in order to come closer to our goal of 64 eligible participants. After additional recruitment and then exclusion, there were 52 participants in our sample, including 10 Black women, 10 Black men, 15 white women and 17 white men in the sample (52 participants total, mean age = 30.71 [SD = 10.21], range = 18-55).

2.5.2. Results

We analysed if perceptions of agency were modulated by conventional social hierarchies or by social similarities between the participant and the depicted figures. This analysis is comprised of two parts, firstly looking at the effect of gender on agency perceptions. To do so, we measured participants’ agency ratings of two figures interacting who differed in gender. Secondly, we looked at the effect of race on agency perceptions, measuring participant’s ratings of two figures interacting who differed in race. Agency ratings were based on the properties of causation and volition.

2.5.2.1. Gender effects

Firstly, we analysed if participants perceived the man as more agentive than the woman on gender trials, and if this was modulated by participant gender, participant race or the type of dyad. We conducted three linear mixed effects models; the first
combining ratings across both causation and volition trials, the second taking a subset of
the data using causation trials only, and the third taking a subset using volition trials
only. The combined model was of the form Agency Rating ~ Participant Gender *
Participant Race * Dyad Race + (1+ Dyad Race | Subject) + (1| Event). The random
effect of dyad was removed from the causation model. The volition model converged
with the random effect of dyad, and the random effect by item of participant gender.
Results are displayed in Table 14.

A non-significant intercept in all three models indicated that there was no
overall bias towards seeing the man as more agentive than the woman (trials combined
M = 1.36 [SD = 6.56], causation-only M = 1.61 [SD = 6.75], volition-only M = 1.12
[SD = 7.61]). However, a significant main effect of dyad race in all three models
suggested that participants were more likely to rate the man as the more agentive person
on trials depicting white woman/white man dyads than Black woman/Black man dyads
(combined: M = 7.64 [SD = 9.97] vs. M = -4.88 [SD = 8.49]). These trends are depicted
in Figure 17.
### Table 14

*Results from the male-agent logistic regressions on gender trials*

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.24</td>
<td>2.67</td>
<td>10.21</td>
<td>0.46</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Participant gender</td>
<td>-1.44</td>
<td>0.89</td>
<td>49.88</td>
<td>-1.63</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Participant race</td>
<td>0.84</td>
<td>0.89</td>
<td>49.88</td>
<td>0.94</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Dyad race</td>
<td>5.90</td>
<td>0.91</td>
<td>50.27</td>
<td>6.50</td>
<td>0.00 ***</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>1.70</td>
<td>0.89</td>
<td>49.88</td>
<td>1.92</td>
<td>0.06 *</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Dyad race</td>
<td>-0.53</td>
<td>0.91</td>
<td>50.28</td>
<td>-0.59</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>Participant race : Dyad race</td>
<td>1.66</td>
<td>0.91</td>
<td>50.27</td>
<td>1.83</td>
<td>0.07 *</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race : Dyad race</td>
<td>0.32</td>
<td>0.91</td>
<td>50.27</td>
<td>0.36</td>
<td>0.72</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Causation-only</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.50</td>
<td>2.83</td>
<td>9.09</td>
<td>0.53</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Participant gender</td>
<td>-1.91</td>
<td>0.88</td>
<td>49.00</td>
<td>-2.17</td>
<td>0.04 *</td>
<td></td>
</tr>
<tr>
<td>Participant race</td>
<td>0.68</td>
<td>0.88</td>
<td>49.00</td>
<td>0.77</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Dyad race</td>
<td>7.03</td>
<td>0.88</td>
<td>989.63</td>
<td>8.01</td>
<td>0.00 ***</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>2.06</td>
<td>0.88</td>
<td>49.02</td>
<td>2.34</td>
<td>0.02 *</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Dyad race</td>
<td>-0.43</td>
<td>0.88</td>
<td>989.66</td>
<td>-0.49</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>Participant race : Dyad race</td>
<td>1.70</td>
<td>0.88</td>
<td>989.67</td>
<td>1.94</td>
<td>0.05 *</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race : Dyad race</td>
<td>-0.29</td>
<td>0.88</td>
<td>989.65</td>
<td>-0.34</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Volition-only</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.97</td>
<td>2.47</td>
<td>10.35</td>
<td>0.39</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Participant gender</td>
<td>-1.02</td>
<td>1.15</td>
<td>15.45</td>
<td>-0.89</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Participant race</td>
<td>1.05</td>
<td>1.07</td>
<td>50.40</td>
<td>0.98</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Dyad race</td>
<td>4.82</td>
<td>0.94</td>
<td>50.60</td>
<td>5.14</td>
<td>0.00 ***</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>1.36</td>
<td>1.07</td>
<td>50.40</td>
<td>1.27</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Dyad race</td>
<td>-0.67</td>
<td>0.94</td>
<td>50.60</td>
<td>-0.71</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Participant race : Dyad race</td>
<td>1.62</td>
<td>0.94</td>
<td>50.58</td>
<td>1.73</td>
<td>0.09 *</td>
<td></td>
</tr>
<tr>
<td>Participant gender : Participant race : Dyad race</td>
<td>0.98</td>
<td>0.94</td>
<td>50.59</td>
<td>1.05</td>
<td>0.30</td>
<td></td>
</tr>
</tbody>
</table>

* *p < .05. **p < .01. ***p < .001. * = marginally significant.*
The mean proportion of trials demonstrating a male-agent bias

Figure 17

Note. The mean agency ratings (causation and volition combined) on gender trials across participant demographic and dyad race. Zero represents chance, above zero represents the man favoured as the more likely agent, below represents the woman favoured as the more likely agent. Error bars represent standard error.

The causation model further showed a significant main effect of participant gender (marginal in the combined model): male participants were more likely than women to rate the man as more agentive than the woman (M = 3.08 [SD = 6.94] vs. M = 0.13 [SD = 6.34]). This model also showed a significant interaction between participant gender and participant race; all demographic groups tended to rate the man as more agentive than the woman, apart from Black female participants who tended to rate the woman as more agentive (see Figure 18).
In all three models, a marginal interaction emerged between participant race and dyad race. There was no effect of adding the fixed effect of participant age to the main models.⁶

**Figure 18**

*The mean proportion of causation trials demonstrating a male-agent bias*

Note. The mean causative agency ratings on gender trials across participant gender and race. Zero represents chance, above zero represents the man favoured as the more likely agent, below represents the woman favoured as the more likely agent. Error bars represent standard error.

⁶Significance values for the models with participant age added as a predictor: combined $p = .25$, causation $p = .18$, volition $p = .41$. 

2.5.2.2. *Race effects*

Next, we analysed if participants perceived the white figure as more agentive than the Black figure on race trials, and if this was modulated by participant race, participant gender and the gender of the dyads. Again, we conducted three linear mixed effects models to analyse the data with the causation and volition trials combined, and then to analyse the data from these trials separately. The combined model was of the form $\text{Agency Rating} \sim \text{Participant Race} \times \text{Participant Gender} \times \text{Dyad Gender} + (1 + \text{Dyad Gender} | \text{Subject}) + (1 + \text{Participant Race} + \text{Participant Gender} | \text{Event})$. In the causation and volition models, the random effect of dyad was removed. The volition model also converged with random effects by item of participant race and participant gender. Results from these models are displayed in Table 15.
Table 15

Results from the white-agent logistic regressions on gender trials

<table>
<thead>
<tr>
<th>Combined</th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.71</td>
<td>2.38</td>
<td>11.14</td>
<td>-1.14</td>
<td>0.28</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.36</td>
<td>1.13</td>
<td>22.79</td>
<td>-0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-2.51</td>
<td>1.18</td>
<td>22.29</td>
<td>-2.13</td>
<td>0.04*</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.39</td>
<td>0.68</td>
<td>49.87</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>1.14</td>
<td>0.99</td>
<td>50.12</td>
<td>1.15</td>
<td>0.26</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>0.68</td>
<td>0.68</td>
<td>49.87</td>
<td>1.00</td>
<td>0.32</td>
</tr>
<tr>
<td>Participant gender : Dyad gender</td>
<td>2.39</td>
<td>0.68</td>
<td>49.87</td>
<td>3.50</td>
<td>0.00**</td>
</tr>
<tr>
<td>Participant race : Participant gender : Dyad gender</td>
<td>0.23</td>
<td>0.68</td>
<td>49.86</td>
<td>0.34</td>
<td>0.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causation-only</th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.06</td>
<td>2.44</td>
<td>9.44</td>
<td>-0.84</td>
<td>0.42</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.82</td>
<td>0.98</td>
<td>50.74</td>
<td>-0.84</td>
<td>0.40</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-3.62</td>
<td>0.98</td>
<td>50.75</td>
<td>-3.70</td>
<td>0.00***</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.65</td>
<td>0.91</td>
<td>989.09</td>
<td>0.72</td>
<td>0.47</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>1.95</td>
<td>0.98</td>
<td>50.75</td>
<td>2.00</td>
<td>0.05*</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>1.21</td>
<td>0.91</td>
<td>989.03</td>
<td>1.33</td>
<td>0.18</td>
</tr>
<tr>
<td>Participant gender : Dyad gender</td>
<td>2.39</td>
<td>0.91</td>
<td>989.00</td>
<td>2.62</td>
<td>0.01**</td>
</tr>
<tr>
<td>Participant race : Participant gender : Dyad gender</td>
<td>0.58</td>
<td>0.91</td>
<td>989.00</td>
<td>0.63</td>
<td>0.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Volition-only</th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.38</td>
<td>2.43</td>
<td>10.73</td>
<td>-1.39</td>
<td>0.19</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.10</td>
<td>1.16</td>
<td>16.81</td>
<td>0.08</td>
<td>0.93</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-1.35</td>
<td>1.23</td>
<td>17.11</td>
<td>-1.10</td>
<td>0.29</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.13</td>
<td>0.87</td>
<td>981.00</td>
<td>0.15</td>
<td>0.88</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>0.30</td>
<td>1.14</td>
<td>50.00</td>
<td>0.26</td>
<td>0.79</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>0.16</td>
<td>0.87</td>
<td>980.85</td>
<td>0.19</td>
<td>0.85</td>
</tr>
<tr>
<td>Participant gender : Dyad gender</td>
<td>2.37</td>
<td>0.87</td>
<td>980.87</td>
<td>2.71</td>
<td>0.01**</td>
</tr>
<tr>
<td>Participant race : Participant gender : Dyad gender</td>
<td>-0.08</td>
<td>0.87</td>
<td>981.03</td>
<td>-0.09</td>
<td>0.93</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001. * = marginally significant.
Figure 19 indicates that participants showed a slight tendency to rate the Black figure as more agentive than the white figure, however non-significant intercepts for each model indicated that this behaviour was not above chance (combined model: $M = -2.80$ [SD = 7.30], causation model: $M = -2.30$ [SD = 7.73], volition model: $M = -3.30$ [SD = 8.01]).

Figure 19

*The mean proportion of trials demonstrating a white-agent bias*

*Note.* The mean agency ratings (causation and volition combined) on race trials across participant demographic and dyad race. Zero represents chance, above zero represents the white person favoured as the more likely agent, below represents the Black person favoured as the more likely agent. Error bars represent standard error.
We found a significant main effect of participant gender in the combined model and the causation model; female participants were less likely than male participants to rate the white person as more agentive than the Black person (combined: $M = -5.04$ [SD = 7.74]) vs $M = -0.56$ [SD = 6.19]). Participant gender was non-significant in the volition model.

We also found a significant interaction between participant gender and dyad gender in all three models (see Figure 12). Men showed a tendency to rate the white person as more agentive than the Black person on trials depicting two women (combined: $M = 1.35$ [SD = 8.84]) and were more likely to rate the Black person as more agentive than the white person on trials depicting two men ($M = -2.45$ [SD = 8.05]). Female participants, on the other hand, generally rated the Black person as more agentive than the white person on both dyad types but were more likely to do so on trials depicting two women (combined: $M = -8.04$ [SD = 9.44]) than on trials depicting two men ($M = -2.13$ [SD = 7.80]).

In the causation model, we also found a marginally significant interaction between participant race and participant gender. We did not find a statistical difference between Black and white participants in either model. There was no effect of adding the fixed effect of participant age to the main models.\(^7\)

### 2.5.3. Discussion

The results of this experiment did not indicate overall biases to rate men as more agentive than women, or white people as more agentive than Black people. Instead, they revealed differing perceptions of agency depending on the properties measured.

\(^7\) Significance values for the models with participant age added as a predictor: combined $p = .58$, causation $p = .45$, volition $p = .73$.)
(causation or volition), and the participant demographic. When rating figures of
differing genders for causation, male participants tended to rate the man as more likely
to have caused an event than the woman, whereas female participants were close to
chance. Ratings of volition did not differ depending on participant gender.

On gender trials, a significant main effect of dyad indicated that participants
were also more likely to rate the man as more agentive than the woman, for both
causation and volition, if both people were white rather than Black. In fact, Figure 17
also indicates that participants were generally more likely to rate the woman as more
agentive than the man if both people were Black. Thus, these ratings suggests that
participants were most likely to perceive white men and Black women as the more
likely agents of the events.

This effect of dyad race on gendered perceptions of agency reflects the findings
of Livingston and colleague's leadership study (2012), who found comparable
perceptions of leadership between white men and Black women. In our study,
participants were rating white men as more agentive than white women, and Black
women as more agentive than Black men; the latter result perhaps echoing the “Strong
Black Woman” trope that expects a degree of agency and leadership from Black women
(Nelson et al., 2016) but ultimately punishes them when their behaviour contradicts this
stereotype (Livingston et al., 2012).

We did not find evidence that white participants were more likely to rate the
white figure as more agentive than Black participants. A main effect of participant
gender suggested that female participants were less likely than male participants to rate
the white person as more causally agentive than the Black person, however we did not
predict an effect of participant gender on race trials and therefore more research would
be required to determine if this effect, and the interaction between participant gender and dyad, is both replicable and rooted in theory.

In summary, perceptions of agency as they relate to causation and volition are unlikely explanations for the Like Me race bias found in our first production experiments. Agency appears to be a potential explanation for the Like Me gender bias, but only insofar as to possibly account for why men showed a stronger Like Me gender bias than women, and only for the agentive property of causation.

While perceived agency of the figures has not offered a convincing explanation for the Like Me bias, we will continue to explore agency, this time as it relates to thematic role. So far, Experiments 1 and 2 have not provided general evidence that differences in grammatical function would affect a speaker’s likelihood to mention the person more like themselves first, apart from the main effect of syntax on the Like Me gender bias in Experiment 1, which we did not replicate in Experiment 2. Our investigations now turn to asking if distinctions between event agent and event patient would have an effect.

Unlike in active transitive sentences, such as those constructed in the previous experiments, constructing transitive passive sentences requires the speaker to make two important but different decisions: 1) who is the agent of the event, which differs from 2) who they will mention first as the sentence subject. Conversely, the first person a speaker mentions while constructing an active sentence is both the agent of the event and the sentence subject. This may present distinctions for who speakers choose to mention first, when describing interactions between dyads of different genders, or of difference races.
To test whether attribution of sentence subject or event agency is a greater determinant of mention order, we ran a further experiment that was almost identical to the first two but included a new syntax condition, this time comparing active and passive sentences. In doing so, we tested whether participants are biased to make the figures more like them the thematic agents, or whether they are biased more towards putting the figures more like them first in the sentence, as the sentence subject.

One possibility is that the Like Me bias behaves like the animacy bias. The animacy bias is argued to occur due to conceptual accessibility, as more animate entities are more accessible to a speaker than less animate entities, making them easier to retrieve from memory. Consequently, the speaker is more likely to attribute subjecthood to the more accessible entity and to mention it first (Bock, 1986; Mcdonald et al., 1993) regardless of grammatical function, as determined by studies that have found speakers show an animacy bias for transitive sentences and conjunctions (Branigan et al., 2008; Branigan & Feleki, 1999; Franz et al., 2021). Speakers also tend to use the passive form when there is a variation of animacy, using passive syntax to ensure the more animate entity is attributed sentence subjecthood over the less animate entity (Prat-Sala et al., 2000; see also Ferreira, 2021), for example, “the postman was chased by the dog”.

So far, we have found the Like Me bias resembles the animacy bias for mention-order preferences, as we found this effect in both transitive and conjoined sentences. If the Like Me bias generally behaves like the animacy bias, then serial order should be the main determinant of first mention, rather than distinctions of grammatical or thematic role. Consequently, as with the animacy bias, the Like Me bias should manifest in both active and passive sentences. We predict that speakers will designate the figure most like themselves as the sentence subject rather than as the thematic agent, thereby mentioning them first in both types of sentences.
What is harder to predict is whether this active/passive manipulation would result in an effect of syntax, with speakers more or less likely to show a Like Me bias on passive trials, or, as in Experiments 1 and 2, to show no difference between the two syntax types. Syntax did not reliably influence the Like Me biases in Experiments 1 and 2, despite conjoined noun phrases requiring the speaker to hold both names in working memory for less time than when constructing active sentences (Tanaka et al., 2005). Therefore, the relative ease of retrieval and production has so far not predicted mention order. Thus, with the evidence we have so far, we may not expect an effect of syntax in this next experiment. What we can expect is that by testing this effect, we will be more confidently able to argue whether it is serial order, subjecthood or event agency that influences who a speaker will mention first.

Furthermore, we have only tested the Like Me bias on a mixed sample of English speakers, including those residing in the UK, Ireland, the United States of America, Canada and Australia. While doing so has allowed these previous studies to be unconstrained by the specificity of one location, thereby making conclusions generalisable across different English-speaking populations, it is important to recognise that culture often has a significant impact on both language and identity. For this reason, the next experiment used a more specific population sample, namely from the USA, controlling for the country of the participant sample and how long the participants have been immersed in that country.

Additionally, we returned to Leach et al.’s Multicomponent In-group Identification Survey in order to investigate how identifying with one’s own gender or race may influence the likelihood that a person would demonstrate a Like Me gender or race bias. We hoped to replicate the findings of Experiment 1, which found that
participants who were more likely to self-stereotype in line with their racial group were more likely to show a Like Me race bias.

Therefore, Experiment 4 sought to replicate the main findings of Experiments 1 and 2, but with a new syntax condition comparing active and passive sentences, and with a population sampled from the USA.

2.6. Experiment 4

Experiment 4 set out to replicate the main findings of Experiment 1 and 2, again testing the effects of conventional social hierarchy and the Like Me gender and race biases, and to build on those experiments by testing a different syntax condition, comparing active and passive sentences. We preregistered the experiment on osf.io (https://osf.io/drb6f).

If using the passive voice influences participants to mention the figure of their own gender first, we would expect to find a significant interaction between syntax and participant gender when testing a man-first bias and a significant effect of syntax on the Like Me gender bias. Likewise, if the passive voice influences participants to mention the figure of their own race first, we would expect a significant interaction between syntax and participant race when testing a white-first bias and a significant effect of syntax on the Like Me race bias.

Additionally, the current experiment tested a more localised participant pool, those who reported having been living in the USA for the past five years, and once again measured in-group identification in a post-study questionnaire. By doing so, we aimed to replicate the findings from Experiment 1: null effects of gender in-group
identification on the Like Me gender bias and a significant effect of self-stereotyping on the Like Me race bias.

2.6.1. **Method**

Methods were as Experiment 1, unless specified otherwise.

2.6.1.1. **Participants**

240 new participants were recruited from Prolific.com (60 white men, 60 Black men, 60 white women, 60 Black women); all English-speakers who had been living in the USA for the last five years.

2.6.1.2. **Materials**

In this experiment, we changed the syntax condition from transitive/conjoined, to active transitive/passive transitive. As in the previous experiments, on half of the trials, participants were prompted to describe the photographs (which were the same as in Experiment 2 and 3) using a fragment of an active transitive sentence (e.g., shown ...is shouting at...). On the other half of the trials, they were prompted using a passive transitive fragment (e.g., ...is being shouted at by...).

Some of the verbs used in the previous experiments would not sound natural in passive voice (e.g., *play, fight*). Thus, we replaced a subset of the verbs: for three verbs, the replacements had similar semantics (*seduce* over *flirt; challenge* over *argue; oppose* over *fight*), in three cases we used new actions to accompany new verbs (replacing *dance, play* and *talk* with *hug; greet; look*), and in one case, we used the same action but with a different verb (the action for *marry* became associated with the verb *copy*). As in the previous experiments, we categorised the verbs into three valence conditions:
intimate (hug, seduce, kiss), casual (greet, look, touch, copy) and negative (challenge, oppose, shout).

We included the same post-study demographics questionnaire as in the previous experiments, with the addition of the question “have you been living in the USA for the last five years?”, to which participants answered yes or no. As in Experiment 1, we measured in-group identification using the two-survey scale developed by Leach et al. (2008). In this study, responses to the scale were given on a 0 – 100 sliding scale (Strongly disagree – Strongly agree), to allow finer gradations of measurement than is possible with a 7-point Likert scale and to reduce the likelihood that participants would select the exact same point on each trial. We included 24 filler questions (14 among the gender-related survey and 14 among the race-related survey) to reduce participant inattention.

Analyses were conducted using the same principles as in Experiments 1 and 2. As before, we analysed if mention order was affected by fixed social hierarchies of gender and race, and if participants tended to show a Like Me bias to mention the person of their own gender first, and the person of their own race first. We first analysed gender trials and then race trials.

2.6.1.3. Exclusions

Exclusion criteria were all pre-registered. We used the same criteria as in Experiment 1 for individual participants and individual trials. We excluded 2 participants whose reported race did not match across the two records and 6 participants who did not report being L1 English speakers. Additionally, we excluded 5 participants who reported not living in the USA for the last five years.
11% of total trials were excluded for response errors and 29 participants were excluded for making response errors on 50% more of active trials, or 50% or more of passive trials. We excluded 5 participants who mentioned the left figure first on 90% of trials and 3 participants who mentioned the right figure first on 90% of trials; mean left-bias score was 51% [SD=0.18]). 2% of trials were excluded for reactions times greater than 3 standard deviations from the participant mean. After exclusions, the main analyses included 196 participants: 46 Black Women, 49 Black Men, 55 White Women and 46 White Men; mean age = 29.97 [SD = 8.78], range 18 - 68).

For the gender in-group identification analysis, one item was removed from the dataset during a reliability check for having an item-total correlation of under 0.50, as per the pre-registered exclusion criteria; item “often” from the Centrality sub-scale (0.47).

2.6.2. Results

This analysis follows the same structure as in Experiments 1 and 2: mention order as affected by androcentrism (man-first) and then a Like Me gender bias, followed by mention order as affected by a white default (white-first) and then a Like Me race bias.

2.6.2.1. Man-first

We analysed if participants showed a tendency to mention the man first on gender trials, and the extent to which mention order was influenced by participant demographic, the syntax of the completed sentence, the race of the dyads and the valence of the event (see Figure 20 for effects on the man-first bias). To do so, we conducted a mixed effects logistic regression of the form Man First ~ (Participant
Gender + Syntax + Participant Race + Dyad Race + Valence)^2 + (1 + Syntax + Dyad Race | Subject). By-item random effects were removed for model convergence. The results from this regression are displayed in Table 16.

**Figure 20**

*The mean proportion of gender trials on which the man was mentioned first*

*Note.* Means for both male and female participants and on active and passive sentences. Data is also distinguished by dyad race (Black Dyads = Black woman/Black man dyads, White Dyads = white woman/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the man, below it represents a bias in favour of the woman. Error bars represent standard error.
Table 16

Results from man-first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.04</td>
<td>0.04</td>
<td>0.88</td>
<td>0.38</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.28</td>
<td>0.04</td>
<td>-6.22</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.03</td>
<td>0.06</td>
<td>0.52</td>
<td>0.60</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.06</td>
<td>0.04</td>
<td>-1.23</td>
<td>0.22</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.03</td>
<td>0.03</td>
<td>1.03</td>
<td>0.30</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.16</td>
<td>0.88</td>
</tr>
<tr>
<td>Participant gender : Syntax</td>
<td>0.00</td>
<td>0.06</td>
<td>0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>Participant gender : Dyad race</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Participant gender : Valence</td>
<td>0.04</td>
<td>0.03</td>
<td>1.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Syntax : Participant race</td>
<td>0.02</td>
<td>0.06</td>
<td>0.30</td>
<td>0.76</td>
</tr>
<tr>
<td>Syntax : Dyad race</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.29</td>
<td>0.77</td>
</tr>
<tr>
<td>Syntax : Valence</td>
<td>-0.08</td>
<td>0.03</td>
<td>-2.43</td>
<td>0.01 **</td>
</tr>
<tr>
<td>Participant race : Dyad race</td>
<td>0.02</td>
<td>0.03</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>Participant race : Valence</td>
<td>-0.03</td>
<td>0.03</td>
<td>-0.91</td>
<td>0.36</td>
</tr>
<tr>
<td>Dyad race : Valence</td>
<td>0.06</td>
<td>0.03</td>
<td>1.68</td>
<td>0.12</td>
</tr>
</tbody>
</table>

**p < .01. ***p < .001.
Unlike our prior studies, this third study – with new stimuli and a new population – did not find an overall bias to mention the man first rather than the woman \( (M = 0.51 \ [SD = 0.17]) \), as indicated by a non-significant intercept for the regression. Critically, we did again find a reliable main effect of participant gender: male participants were significantly more likely than female participants to mention the man first \( (M = 0.57 \ [SD = 0.16] \text{ vs. } M = 0.45 \ [SD = 0.16]) \).

There was a significant interaction between syntax and valence (Figure 21), which suggested that participants were more likely to mention the man first on passive than active sentences, but only for the negative trials \( (\text{passive } M = 0.53 \ [SD = 0.29], \text{ active } M = 0.49 \ [SD = 0.28]) \). We did not find a significant interaction between syntax and gender.
Figure 21

Effects of valence on the man-first bias

Note. The mean proportion of gender trials on which participants demonstrated a man-first bias, on active and passive sentences and across the valence condition. The dotted line represents chance at 50% - above the line represents a bias in favour of the man, below it represents a bias in favour of the woman. Error bars represent standard error.

2.6.2.2. Like Me Gender bias

Next, we investigated if the mention order behaviour of participants was consistent with our Like Me bias hypothesis on gender trials (see Figure 22 for effects on the Like Me gender bias). To do so, we conducted a mixed effects logistic regression of the form Like Me ~ Participant Gender + Syntax + Participant Race + Dyad Race + Valence + (1 + Syntax + Dyad Race | Subject) + (1 | Event), results are displayed in Table 17.
Figure 22

*The mean proportion of gender trials demonstrating a Like Me gender bias*

*Note.* Showing means of participants from all four demographic groups, on active and passive trials and across dyad race. Black Dyads = Black woman/Black man dyads, White Dyads = white woman/white man dyads. The dotted line represents chance at 50% - above the line represents a Like Me gender bias. Error bars represent standard error.
Table 17

Results from the Like Me gender bias linear regressions

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.28</td>
<td>0.05</td>
<td>5.19</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.00</td>
<td>0.06</td>
<td>-0.08</td>
<td>0.94</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.02</td>
<td>0.04</td>
<td>0.40</td>
<td>0.69</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.00</td>
<td>0.03</td>
<td>-0.14</td>
<td>0.89</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.72</td>
<td>0.47</td>
</tr>
</tbody>
</table>

***p < .001.

We found that participants were more likely to mention the figure of their own gender first rather than the figure of a different gender (M = 0.56 [SD = 0.13]), as indicated by a significant intercept, therefore providing evidence that supports our hypothesis of a Like Me gender bias. Unlike in Experiments 1 and 2, we did not find a significant effect of participant gender. However, one-sample t-tests again showed that Black women did not show a significant Like Me gender bias whereas the other groups did (Black women M = 0.53 [SD = 0.17], t(45) = 1.54, p =.07; Black men M = 0.58 [SD = 0.18], t(48) = 3.47, p <.01; white women M = 0.57 [SD = 0.15] , t(54) = 4.06, p <.01; white men M = 0.56 [SD = 0.14], t(45) = 3.83, p <.01). Syntax was not a significant predictor of the Like Me gender bias.
**Participant gender.** To determine if the Like Me gender bias was indeed present in both female and male populations, we conducted a mixed effects logistic regression on both groups independently. These were of the form Like Me ~ Participant Race + Syntax + Dyad Race + Valence + (1 + Syntax + Dyad Race | Subject) + (1 | Event). The results of both models are shown in Table 18.

### Table 18

*Male and female results from the Like Me gender bias linear regressions*

<table>
<thead>
<tr>
<th></th>
<th>Male participants</th>
<th>Female participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.32</td>
<td>0.07</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>Dyad race</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.03</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**$p < .01$. ***$p < .001$.**

For both male and female participants, the intercept was significant, indicating that both groups were reliably showing a Like Me gender bias. No other fixed effects were significant.
2.6.2.3. Gender In-group identification

As in Experiment 1, we next investigated how individual differences may have influenced the Like Me gender bias by analysing the two surveys from Leach et al. (2008), pertaining to gender in-group identification ((Group-Level) Self-Investment and (Group-Level) Self-Definition). We entered participants’ totals on these surveys as scaled fixed effects in the model, of the form Like Me ~ Participant Gender + Syntax + Participant Race + Dyad Race + Valence + Self-Investment + Self-Definition + (1 + Syntax + Dyad Race | Subject) + (1 | Event). As in Experiment 1, neither survey predicted the Like Me gender bias (Self-Investment: \( p = .71 \); Self-Definition: \( p = .98 \)).

Control analyses tested the effects of trial progression and participant age. We did not find an effect of trial progression on the Like Me gender bias, however younger participants were slightly more likely to show a Like Me gender bias than older participants\(^8\). Next, we analysed data from the race trials.

2.6.2.4. White-first

Next, we analysed if participants showed a tendency to mention the white figure first on race trials, and the extent to which mention order was influenced by participant demographic, syntax, the gender of the dyads and event valence (see Figure 23 for effects on the white-first bias). To do so, we conducted a mixed effects logistic regression of the form White First ~ (Participant Race + Syntax + Participant Gender + Dyad Gender + Valence) \(^2\) + (1 + Syntax + Dyad Gender | Subject). By-item random effects were removed for model convergence. Results from this model are shown in Table 19.

---

\(^8\) Progression across the experiment did not affect the likelihood of a Like Me gender bias (\( p = .45 \)). Younger participants were significantly more likely to show a Like Me gender bias on these trials (\( \beta = -0.11, SE = 0.04, z = -2.53, p < .05 \)).
The mean proportion of race trials on which the white person was mentioned first

Note. Means for both white and Black participants and on active and passive sentences. Data is also distinguished by dyad gender (Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads). The dotted line represents chance at 50% - above the line represents a bias in favour of the white person, below it represents a bias in favour of the Black person. Error bars represent standard error.
### Table 19

*Results from the white-first linear regression*

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.81</td>
<td>0.42</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.06</td>
<td>0.04</td>
<td>1.64</td>
<td>0.10</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.07</td>
<td>0.06</td>
<td>-1.27</td>
<td>0.20</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.01</td>
<td>0.03</td>
<td>0.21</td>
<td>0.83</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.06</td>
<td>0.04</td>
<td>1.60</td>
<td>0.11</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.04</td>
<td>0.03</td>
<td>-1.13</td>
<td>0.26</td>
</tr>
<tr>
<td>Participant race : Syntax</td>
<td>0.00</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.97</td>
</tr>
<tr>
<td>Participant race : Dyad gender</td>
<td>0.01</td>
<td>0.03</td>
<td>0.19</td>
<td>0.85</td>
</tr>
<tr>
<td>Participant race : Participant gender</td>
<td>0.01</td>
<td>0.04</td>
<td>0.17</td>
<td>0.86</td>
</tr>
<tr>
<td>Participant race : Valence</td>
<td>0.03</td>
<td>0.03</td>
<td>0.82</td>
<td>0.41</td>
</tr>
<tr>
<td>Syntax : Dyad gender</td>
<td>0.01</td>
<td>0.03</td>
<td>0.22</td>
<td>0.82</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>-0.09</td>
<td>0.06</td>
<td>-1.62</td>
<td>0.10</td>
</tr>
<tr>
<td>Syntax : Valence</td>
<td>0.05</td>
<td>0.03</td>
<td>1.44</td>
<td>0.15</td>
</tr>
<tr>
<td>Dyad gender: Participant gender</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.38</td>
<td>0.70</td>
</tr>
<tr>
<td>Dyad gender : Valence</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.23</td>
<td>0.81</td>
</tr>
<tr>
<td>Participant gender : Valence</td>
<td>0.01</td>
<td>0.03</td>
<td>0.34</td>
<td>0.73</td>
</tr>
</tbody>
</table>

A non-significant intercept indicated that participants were not reliably more likely to mention the white figure first, rather than the Black figure (M = 0.49 [SD = 0.11]). Unlike Experiments 1 and 2, the main effect of participant race was not significant.
2.6.2.5. *Like Me Race bias*

Next, we reanalysed these data in terms of whether participants were more likely to firstly mention the person of their own race (see Figure 24 for effects on the Like Me race bias). For this, we used a mixed effects logistic regression of the form \( \text{Like Me} \sim \text{Participant Race} + \text{Syntax} + \text{Participant Gender} + \text{Dyad Gender} + \text{Valence} + (1 + \text{Syntax} + \text{Dyad Gender} | \text{Subject}) \). Random effects by item were removed for model convergence. The results of this model are displayed in Table 20.

**Figure 24**

*The mean proportion of race trials demonstrating a Like Me race bias*

![Figure 24](image)

*Note.* Showing means of participants from all four demographic groups, on active and passive trials and across dyad gender. Female Dyads = Black woman/white woman dyads, Male Dyads = Black man/white man dyads. The dotted line represents chance at 50% - above the line represents Like Me race bias. Error bars represent standard error.
A marginally significant intercept indicated that participants were slightly more likely than chance to mention the figure of their own race first (M = 0.51 [SD = 0.11]). No other fixed effects were significant.

**Participant race.** To determine if the Like Me race bias was present in both Black and white populations, we conducted a mixed effects logistic regressions on both groups independently. For both groups, these models were of the form Like Me ~ Syntax + Participant Race + Participant Gender + Dyad + Syntax + Valence + (1 + Dyad + Syntax | Subject). Random effects by item were removed for model convergence. Results from these models are displayed in Table 21.
Table 21

Results from the Like Me race bias linear regressions for Black and white participants

<table>
<thead>
<tr>
<th></th>
<th>Black participants</th>
<th>White participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Valence</td>
<td>0.07</td>
<td>0.05</td>
</tr>
</tbody>
</table>

*= marginally significant.

For Black participants, the intercept was marginally significant, with no other fixed effect presenting as significant. For white participants, neither the intercept nor any of the fixed effects were statistically significant.

2.6.2.6. **Racial In-group identification.**

Next, we investigated how individual differences may have influenced the Like Me race bias. We entered participants’ scores from the in-group identification subscales for race, each as a scaled fixed effect, into a new model of the form Like Me ~ Participant Race + Syntax + Participant Gender + Dyad Gender + Valence + Self-Investment + Self-Definition + (1 + Syntax + Dyad Gender | Subject). Random effects by item were removed for model convergence. Results from this model are displayed in Table 22.
Table 22

*Results from the Like Me race bias regression with surveys as added fixed effects*

<table>
<thead>
<tr>
<th>Estimate</th>
<th>( \beta )</th>
<th>SE</th>
<th>( z )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.06</td>
<td>0.04</td>
<td>1.82</td>
<td>0.07 *</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.10</td>
<td>0.05</td>
<td>-2.15</td>
<td>0.03 *</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.01</td>
<td>0.06</td>
<td>-0.16</td>
<td>0.87</td>
</tr>
<tr>
<td>Dyad gender</td>
<td>0.01</td>
<td>0.03</td>
<td>0.19</td>
<td>0.85</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.89</td>
</tr>
<tr>
<td>Valence</td>
<td>0.03</td>
<td>0.03</td>
<td>0.82</td>
<td>0.41</td>
</tr>
<tr>
<td>Self-investment</td>
<td>-0.16</td>
<td>0.06</td>
<td>-2.53</td>
<td>0.01 *</td>
</tr>
<tr>
<td>Self-definition</td>
<td>0.08</td>
<td>0.05</td>
<td>1.61</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*\( p < .05 \). **\( p < .01 \). ***\( p < .001 \). * = marginally significant.*

We found a significant predictor in (Group-level) Self-investment; participants who scored higher on this survey were less likely to demonstrate a Like Me race bias, contrary to our predictions. Results again showed a marginally significant intercept. However, this model also showed a significant main effect of participant race; Black participants showed a stronger Like Me race bias than white participants (M = 0.52 [SD = 0.14] vs M = 0.51 [SD = 0.14]). Random effects by item were removed for model convergence, however when participant race was included as a random effect, it still showed significance, as did (Group-level) self-investment.
To understand this in-group identification measure further, we ran an additional model with participants scores on all five survey items added as predictors, however none of these items were significant predictors of the Like Me race bias. Control analyses tested the effects of trial progression and participant age. We did not find an effect of either on the Like Me race bias.\footnote{Trial progression was not a significant predictor of the Like Me race bias ($p = .27$), nor was participant age ($p = .10$).}

### 2.6.2.7. Memory effects

Finally, we explored possible memory effects on this task. We examined participants’ accuracy at remembering the names of the eight photographed figures, analysing mean trial success across participants during the first round of the memory task. Overall, participants were demonstrating relatively high accuracy on this task (Black women $M = 0.95$ [SD = 0.14]; Black men $M = 0.95$ [SD = 0.13]; white women $M = 0.96$ [SD = 0.12]; white men $M = 0.93$ [SD = 0.15]). To determine if participants were better at memorising certain stimuli over others, we used a mixed effects logistics regression of the form Accuracy ~ Participant Gender * Stimulus Gender + Participant Race * Stimulus Race + (1 | Subject). Random effects by item were removed for model convergence, as were non-significant interactions.

We found a significant effect of stimulus gender on memory task accuracy; participants were better at remembering names for men than women ($\beta = -0.33$, $SE = 0.10$, $z = -3.30$, $p < .001$. This effect was marginal in Experiment 1 and significant in Experiment 2.
2.6.3. Discussion

We tested if gender affected mention order biases and found that participants were not overall more likely to mention the man first and the woman second, but that they were more likely to mention the figure of their own gender first. This did not vary by participant gender; female and male participants were as likely to show a Like Me gender bias. Out of the four demographic groups tested, Black women were the least likely group to demonstrate this bias.

We also tested if race affected mention order biases and found that participants were not overall more likely to mention the white person first and the Black person second. However, participants showed a marginal tendency to mention the figure of their own race first, over the figure of a different race. On inspection of all four demographic groups, descriptive statistics and one-sample t-tests indicated that only Black men demonstrated a significant Like Me race bias on these trials (M = 0.53 [SD = 0.15], t(48) = 1.82, p < .05; vs. white women M = 0.52 [SD = 0.14], t(54) = 1.23, p = .11; Black women M = 0.50 [SD = 0.13], t(45) = 0.27, p = .39; white men M = 0.49 [SD = 0.15], t(45) = -0.42, p = .66).

As per our predictions, participants showed a Like Me bias on both active and passive sentences, however this did not differ by syntax. The type of dyad also had no effect on mention order (whether the dyads were Black woman/Black man pairs or white woman/white man pairs on gender trials, or Black woman/white woman pairs or Black man/white man pairs on race trials).

As in Experiment 1, none of our in-group identification measures (using surveys from Leach et al., 2008) predicted the Like Me gender bias. However, we found that the Like Me race bias decreased in likelihood the higher participants scored on the Self-Investment survey. This is the opposite direction to what we predicted and does not
replicate the null effect found in Experiment 1. This result indicates that more research is required to understand the interplay between social cognition, group identification and language choices. Lastly, we analysed participants’ accuracy on the first round of the memory game and, again, found that they were better at remembering the names of men than of women.

2.7. **General Discussion**

Previous literature has provided good evidence that mention order is not arbitrary and that there are, in fact, reliable social factors that influence who a speaker will mention first and who they will mention second when producing language. We carried out four experiments to test this notion, focusing on effects of gender and race on first mention.

2.7.1. **Support for main hypotheses**

In each production experiment, we tested the effects of conventional social hierarchies on mention order. When investigating effects of gender, we found evidence for a man-first bias in mention order in both Experiments 1 and 2, but not in Experiment 4. When investigating effects of race, we did not find a white-first bias in either Experiment. This suggests that conventional social hierarchies affect mention order as it relates to gender to some extent, but not as it relates to race.

We also tested our Like Me hypothesis, that speakers will show a tendency to mention a person more like themselves first and a person less like themselves second, and this manifested as the more compelling explanation for the mention order behaviours observed in this study. We found a significant Like Me gender and Like Me
race bias across all three production experiments, supporting the discovery of a newly conceptualised production behaviour. Men were more likely than women to demonstrate a Like Me gender bias in Experiments 1 and 2, which suggests that mention order here was likely influenced by both a Like Me effect and an effect of androcentrism. In other words, women showed a tendency to mention the woman first and the man second, but this was also modulated by a general tendency to put the man first. However, in Experiment 4 men and women did not differ in their likelihood to mention the person of their own gender first. Paired with the finding that there was no overall man-first bias in that experiment, this result indicated that mention order was influenced more by relative similarity with the speaker rather than effects of androcentrism.

We also found evidence for a Like Me race bias each time, and this did not differ based on participant race in either of these three production experiments. Although the Like Me race effect was marginal in Experiment 4, it was in the same direction as in Experiments 1 and 2, which points toward a reliable trend.

Why was this effect weaker in Experiment 4, and why did we find a man-first bias in the first two experiments but not in the final one? One possibility is that this is a population effect. In the first two experiments, the population was sampled from different English-speaking countries, whereas in Experiment 4, the sample population was from America alone. Influences of gender and race on behaviour are likely to be highly influenced by societal norms and cultural attitudes. Therefore, more complex investigations of these effects may want to be careful to consider specific populations at first, before widening their scope.

Another possibility is that the types of events used in these experiments accounts for the differences in results. In Experiments 1 and 2, the ten events used were very
similar, with only two differences in actions and verbs across the stimuli set (hug and meet in Experiment 1 and flirt and play in Experiment 2). In Experiment 4, however, the differences between events in this stimuli set and the previous experiments were more pronounced, as it was necessary to present sentences that would sound natural in both active and passive form, so new verbs were needed. It is likely that there are important differences between the verbs that we had chosen to replace the originals, despite their synonymous qualities (i.e., argue versus challenge, fight versus oppose, meet verses greet, and flirt versus seduce). Furthermore, Experiment 4 was the only one to use the verb copy and the verb and action for look. Replication would help determine the extent to which stimuli effects accounted for the variation in main findings between the first experiments and Experiment 4.

2.7.2. **Effects of syntactic form**

Our exploration of syntax reveals important elements of a speaker’s decision-making when describing dyadic interactions. In Experiments 1 and 2, we compared mention order across transitive and conjoined sentences, to investigate the influence of thematic role attribution on mention order. If distinctions of sentence subject and direct object are an important factor of mention order decisions, we would have expected participants to show stronger biases on transitive trials, compared to conjoined trials where both people were the sentence subject. However, this pattern only manifested in Experiment 1 for the Like Me gender bias. It did not replicate in Experiment 2, nor did we find differences between transitive or conjoined trials for the Like Me race bias.

In Experiment 4, we compared mention order across active and passive sentences, to further understand the syntactic effects. If the attribution of thematic agent drives the Like Me bias, we would expect speakers to prescribe event agency to the
person who is most socially similar to themselves. This would manifest as biases of first mention on active sentences and of second mention on passive sentences. Instead, we found that participants were just as likely to show a Like Me bias of first mention when producing passive sentences as when they were producing active sentences, mentioning the person of their own gender or of their own race first and the person who was less similar to themselves second. This result suggests that the attribution of agency is an unlikely explanation for the Like Me bias. Taken together with the findings of Experiment 3, the results from this study do not suggest that perceptions of agency explain the Like me bias.

Instead, the results from investigations of syntactic form in Experiments 1, 2 and 4 suggest that serial order modulates first mention. Speakers tended to mention the person most like themselves first regardless of whether there were distinctions of grammatical function (as in transitive sentences) or not (as in conjoined sentences) and regardless of who would be designated the thematic agent, even if this would be the person least like themselves (as in passive sentences). These findings also reflect those which have shown that a “Me First” bias occurs on binomials where there are no distinctions of grammatical function (Cooper & Ross, 1975; Hegarty et al., 2011; Oeberst & Matschke, 2017; Tachihara & Goldberg, 2020) and that an animacy bias is detected on noun conjuncts (Franz et al., 2021) and passive sentences (Prat-Sala & Branigan, 2000).

2.7.3. Effects of dyad demographic

This study highlights the importance of replication for drawing conclusions about intricate and unexpected effects, for instance in identifying that the effect of syntactic form in Experiment 1 was, importantly, not reliable. The varying influence of
dyad was another key example of the importance of replication, as we did not find the same effects of dyad race on gender trials and dyad gender on race trial across experiments. In Experiment 1, participants were more likely to mention the man first when he was white rather than Black. However, in Experiment 2, participants were more likely to mention the man first if he was Black rather than white, and there was no effect of dyad race on mention order in Experiment 4. Furthermore, in Experiment 1, the Like Me gender bias was strongest on trials depicting two Black people rather than two white people, but in both Experiments 2 and 4 there was no effect of dyad race on the Like Me gender bias.

Then, in Experiment 1, a significant effect of dyad on the white-first bias indicated that participants were more likely to mention the white figure first when dyads were two men than when they were two women. However, dyad did not influence whether participants would be more likely to mention the white person first in Experiment 2 or 4. Additionally, there was no effect of dyad on the Like Me race bias in Experiments 1 or 4.

2.7.4. Intersectional effects and individual differences

An advantage of this study is the inclusion of four different and specific groups, allowing us to interrogate the effect of both gender and race, in addition to how they intersect. Black women have now been found to be the least likely out of all the four demographic groups tested in this study to show a Like Me gender bias. This repeated finding demands reflection. When we consider that Black women are represented by neither of the socially dominant hierarchies explored in this study, on account of being neither male, nor white, this might make them the least likely social group to be mentioned first if a hierarchical explanation for perspective-taking is accepted (that
being, the more socially dominant group would be the most likely to be mentioned first).

However, our investigations of the race trials implied that mention order was more likely driven by the tendency to mention one’s own race first, rather than the white figure first, contrary to a hierarchical explanation and supportive, instead, of a Like Me bias. Furthermore, while there was some evidence for a man-first bias on gender trials in both Experiment 1 and 2, examinations of the four individual demographic groups revealed that in both experiments, white women were in fact more likely to mention the woman first. Thus, Black men, white men and white women were all demonstrating a Like Me gender bias in both experiments, whereas Black women were not. In Experiment 4, all four groups did show a Like Me gender bias, but the bias for Black women was weaker than the rest.

So far, accounts from person perception research do not offer substantial explanations for why Black women would be the least likely group out of the four to show a Like Me gender bias. Research has indicated that the way in which Black women are perceived as agentive differs from that of Black men and or white women and may even show similarities with white men, however that research has not centred Black women as participants (Livingston et al., 2012), and therefore it is unclear how they would have rated themselves for qualities such as leadership status. Furthermore, as stated, agency perceptions did not appear to drive the Like Me gender bias. This is particularly true in the case of Black female participants in Experiment 3, who rated the Black women as more agentive than the Black men they were interacting with (and were close to chance for ratings of agency between white women and men). However, although they may have perceived the Black women to be more agentive, this has not correlated with Black women mentioning themselves first on gender trials.
Instead, to begin to understand the diverging trend for Black women, we have to look elsewhere, to literature that appreciates that meanings of race differ in their internalisation by Black people and white people, and in turn, that the gendering of Black women and their alignment with womanhood differs from that of white women as a result of historical stereotypes, prejudices, pressures and resistances (Higginbotham, 1992). To expect Black women to behave similarly to Black men would ignore that Blackness is further differentiated by the positionality of gender, and to expect Black women to behave similarly to white women would ignore that commonly held assumptions of gender are not necessarily applicable to distinctions of race (Douglass, 2020).

While Black men and white women and men were readily taking the perspective of the figure of their own gender, Black women were unlikely to show a Like Me gender bias and were even proportionally more likely to choose the male perspective at times. This finding, combined with that which showed Black women as appearing relatively more agentive than Black men and white women across all four demographic groups, calls back to the positioning of Black women in discourse as both visible and invisible, and present and absent (Lewis, 2017).

As for our exploration of individual differences, while inconclusive, the results from our analyses as they relate to in-group identification suggest that there is value in considering the ideologies, attitudes, and associations of the speaker as an individual, rather than just their ostensible identity traits. We did not find reliable evidence that people will be more likely to show a Like Me gender or race bias if they also strongly identified with their gender or race, which we measured using Leach and colleagues’ Multicomponent In-group Identification Survey (2008). Alternative measures of in-
group identification may offer different results, however, for instance an instrument that focuses more specifically on interactions.

2.7.5. Effects of agency

Thus, in-group identification does not appear to explain the Like Me bias. In Experiment 3, we also tested perceptions of agency as a possible explanation for the mention order biases found in the first two experiments. However, we did not find evidence that people were generally more likely to rate a person of their own gender or of their own race as more agentive than a person or a different gender or race. While male participants tended to rate the man as more likely to have caused an event than the woman, female participants did not generally show a bias in their perceptions and there was no difference in ratings between men and women for the volitional agent measure. Thus, if judgements of agency are to be considered an influencing factor of mention order, this finding could possibly help explain why the Like Me gender bias tended to appear stronger for male participants than for female participants in Experiments 1 and 2.

However, it does not offer a strong explanation for the overall man-first bias found in both experiments, or the Like Me gender bias demonstrated by female participants. The finding that Black female participants tended to rate the depicted Black woman as the more likely agent than the Black man further dispels the argument that perceptions of agency from the visual stimuli would have influenced mention order, considering Black women have so far been the least likely demographic group to mention the person of their own gender first.

To test this, it might be worthwhile to consider other properties of a prototypical agent highlighted by Dowty, such as movement and sentience (Dowty, 1991), to explore
the question of mention order and perceptions of agency further. There may also be definitions of agency outwith Dowty’s Proto-roles Hypothesis which might be more effective at capturing perceptual differences between participants in events. Overall, however, our findings do not suggest that a Like Me bias of production is driven by a Like Me bias of agency perception.

2.7.6. Control and exploratory analyses

Our control analyses revealed some important findings. In all three production experiments, participants showed a better ability for remembering the names of the men than the women. Crucially, stimulus gender did not interact with participant gender, indicating that all participants found name of the women harder to remember. As we did not find evidence that participants were better at memorising figures of their own gender, or figures of their own race, memory offers a poor explanation for why participants demonstrated a Like Me bias in their mention order. It is possible that the male names used in this study were easier to remember than the female names. If so, this would suggest that these memory effects could be attributable to design. Future experiments should therefore use different monosyllabic names and perhaps involve a norming study to ensure all names are similarly easy to remember. However, and importantly, they do not suggest that memory should be used as an explanation for the Like Me biases.

We did not find consistent evidence that mention order was affected by trial order, suggesting that the paradigm used in this study, including the number of trials presented, was an appropriate measure of a Like Me bias, nor did we find good evidence that the age of the speaker would affect the extent to which they demonstrated a Like Me bias.
Results from Experiment 1 prompted us to explore effects of event valence more closely. In this first experiment, we found effects of valence on the man-first bias and on the white-first bias. Participants were most likely to mention the man first and the woman second on trials depicting intimate events, compared to casual or negative events, and were most likely to mention the white person first on trials depicting negative events, compared to casual or intimate events. These findings were replicated in Experiment 2 but were not replicated in Experiment 4. Furthermore, we did not find evidence for reliable effects of valence on the Like Me biases. This study was not initially designed to incorporate a robust investigation of valence – with only a few events categorised as intimate, casual or negative, it does not offer a highly powered analysis of this factor. However, valence may be a topic worth exploring more stringently in future studies.

2.7.7. Implications for comprehension

Our research suggests that, rather than considering who is a likely agent, for instance based off societal norms, the way in which a speaker conceptualises these interactions and chooses to convey a meaning about them is inherently related to the speaker as an individual. A pressing issue, therefore, is what this means for interpretation. There are many instances in which perspectives are relied on, from literature to testimonies in courts. Whoever is giving their account of an event is making a choice about whose eyes the listener will see that event through. As we have revealed that societal biases do exist in mention order and perspective taking, further research should begin to analyse the effects on the listener (or reader, in the case of the written word), to determine the extent to which mention order may influence language processing and perceptions of individuals.
This study demonstrates that research investigating mention order for descriptions of social interactions should take the speaker’s complex identities, attitudes, and group identification into account. Through an interdisciplinary approach considering work from psycholinguistics, social cognition and social psychology, we can begin to understand the ways in which a speaker’s identity affects perspective taking, or whose eyes the world is seen through.
3. Study 2: Effects of first mention on agency judgements

3.1. Introduction

What semantic inferences would a listener or speaker make from the sentence “Kate is hugging Beth”, and would these differ from the sentence “Kate and Beth are hugging”? The following chapter seeks to understand how mention order influences judgements of role in an interaction, testing the ways in which first mention modulates perceptions of agency. Is Kate the most likely instigator of the hug in both cases, would a reader be sensitive to the subtle semantic differences between these two examples, or would they simply assume, in both cases, that both people were jointly involved?

This work responds to previous studies that have shown how mention order in language production contains biases. For example, speakers tend to mention humans before non-human entities (e.g., preferring “the hiker ran away from the bear” over “the bear chased the hiker”), a finding known as the animacy bias (McDonald et al., 1993; Prat-Sala & Branigan, 2000). They are also more likely to mention more imageable concepts earlier than more abstract concepts (Bock & Warren, 1985) and more prototypical nouns earlier than less prototypical nouns (Onishi et al., 2008). Research has also found biases of social factors in mention order. For instance, studies have shown that people often name men first and women second in noun pairs joined by a conjunction, known as binomials, (Ahmad & Shah, 2019), that they show a tendency to put the names of people closer to themselves in relationship first over people less close, such as when naming couples (Hegarty et al., 2011; Tachihara et al., 2019; Tachihara & Goldberg, 2020), and that they put their in-group first when naming a conflict involving this group and an outgroup (Oeberst & Matschke, 2017).
In Chapter 2, we provided evidence for a new set of biases. First, we found that people are sometimes biased to make certain demographic groups the subjects of their sentences in a way that reflects conventional social hierarchies. Specifically, we found that people showed a tendency to mention a man first and a woman second when describing symmetrical dyadic interactions (for instance describing an interaction of shaking hands as “Fred is meeting Kate”). More reliably, however, this was qualified by a “Like Me bias” for speakers to mention a figure more like themselves first over a figure less like themselves. We tested gender and race as properties of relative social similarity and found that men tended to mention the man first when describing interactions between a woman and a man, while women tended to mention the woman first, and that Black people tended to mention the Black person first when describing interactions between and Black and a white person, while white people tended to mention the white person first. These biases show how basic psycholinguistic processes can be influenced by social cognition.

Importantly, we found the Like Me gender and race biases manifested in utterances even when mention order did not impact on the word’s grammatical role. The size of the Like Me bias was similar for transitive active sentences, in which the two named entities have distinct grammatical and thematic roles (e.g., “Kate is kissing Beth”) and for conjoined sentences, where the two named entities have the same grammatical and thematic roles (e.g., “Kate and Beth are kissing”). Additionally, speakers also demonstrated a Like Me bias when doing so meant that they were no longer assigned the agent thematic role in the sentence, i.e., on transitive passive sentences where the first mentioned entity is the sentence subject but not the agent (e.g., “Kate is being kissed by Beth”). Thus, across these syntactic variations, speakers were biased to mention the figure of their own gender or their own race first.
This chapter aims to investigate the effects of mention order biases on a comprehender (a listener or a reader), focusing on the types of judgements they might make about roles based on first and second mention. Kesebir (2017) has shown that one effect might be perceptions of relevance. When investigating lexically gendered noun pairs (e.g., mother and father), she found that the first noun will be perceived as more relevant or important in contexts where gender beliefs would bolster this assumption. For example, when participants read “My father and mother met my tennis coach last week”, they perceived the father as more involved in the child’s tennis practice than the mother, but when they read “My mother and father met my tennis coach last week”, they did not then show the same bias to perceive the mother as more involved. This was the case for both female and male participants and was thought to be due to societal assumptions that a dad would be more interested in his child’s athleticism than a mum.

Prior research has also shown that people make semantic inferences from mention order that link subjects with thematic agents (Bowerman, 1990; Dowty, 1991; Foley & Van Valin, 2009; Johnson, 1967; Kako, 2006). Agency, defined by Gray and colleagues (2007) as the ability to plan and act, is conceptually about the actions of an entity in relation to a predicate. The relationship between agency and word order follows from theories of thematic roles; the semantic role that a noun plays in relation to a verb according to its grammatical function. In the example “Kate is kissing Beth”, Kate is mentioned first, and her grammatical function is that of the sentence subject. Because the sentence is in the active voice, this means that she is also the semantic agent (the figure doing the action, e.g., the kisser). Beth, who is mentioned second, takes on the grammatical function of sentence object and is also the semantic patient (the figure to which the action is being done, e.g., the kissed). The semantic agent and
patient therefore differ from each other in how they are involved in the event expressed by the verb.

Dowty (1991) provided suggestions for defining features of prototypical agents, including causation (causing another entity to experience an event or change in state), volition (choosing to be involved in the event or state change), sentience and/or perception, movement (relative to another entity’s position), incremental theme (moving towards, over or into another entity) and independent existence (of the event denoted by the verb). His Proto-Roles hypothesis supposes that speakers will show a tendency to ascribe properties of prototypical agents to a sentence subject over the sentence object, and that they will ascribe properties of prototypical patients (e.g., has state change, causally affected by another, does not exist without event) to a sentence object rather than subject.

Kako (2006) provided empirical evidence for Dowty’s principles by measuring people’s intuitions about the properties of individuals in grammatical roles. Firstly, Dowty’s original phrasing of the prototypical agency properties that could be tested with clear and easy to understand questions, or without negating other properties, was translated into layman’s terms (e.g., “volitional involvement” into “chose to be involved”, “sentience” into “aware of being involved”, “causing an event or change of state” into “caused the [noun] to do something” and “caused a change in the [noun]”). Participants read a series of sentences using transitive verbs (e.g., deny, maintain, help, show) and two nonsense nouns, then answered questions of subjectivity, rating the likelihood of the nouns in the previous sentences for having a specific property (e.g., having read the “dax helped the zarg”, they answered the question “How likely is it that the dax moved?”), rating this likelihood on a 1 – 7 Likert scale. Some participants provided ratings of thematic role properties (e.g., volition, causation, sentience), and
others provided ratings of grammatically irrelevant properties (e.g., can be photographed, can make noise). Results from this study showed that people perceive subjects as more agentive than objects, and that they perceive objects as more patient-like than subjects. Interestingly, results did not support that people make similar judgements for subjects and objects with grammatically irrelevant properties, highlighting the importance of the properties Dowty had originally outlined in his Proto-Roles Hypothesis.

That people perceive differences between subjects and objects – for instance as in “Kate is hugging Beth” – is reasonable, considering the semantics implied by agent/patient thematic roles (Bowerman, 1990; Foley & Van Valin, 2009). Based off Kako’s findings, we would expect people to judge Kate as having qualities of an agent and Beth as having qualities of a patient. Less clear, however, is whether people would make the same or similar assumptions when order of mention is the only differentiating factor between the named individuals, for instance in the case of conjoined sentences such as “Kate and Beth are hugging”. Kesebir’s work provided some indication that distinctions of grammatical or thematic role are not required for readers to make biased interpretations from mention order. In her stimuli, both nouns shared the subject position as they were presented in conjoined sentences and were therefore ascribed the same semantic roles. Regardless, readers still made judgements based on the order or mention (Kesebir, 2017). Furthermore, Oeberst and Matschke (2017) have found that people tend to judge a first-mentioned group as more important or powerful when reading binomical conflict names, such as “the Polish-Russian war”. However, Kesebir was testing judgements of relevance and Oeberst and Matschke were testing judgements of power, not agency, so the effect of mention order on perceptions of agency in conjoined sentences remains unclear.
In the current experiment, we tested how participants will interpret an interaction and the people involved in it, based off the order in which these people are mentioned in the event’s description. We aimed to investigate the extent to which mention order is enough for someone to form judgements about the roles people played in an interaction, predicting that participants would show a tendency to judge the person who was mentioned first as more agentive than the person who was mentioned second. To do so, we tested and compared transitive sentences (e.g., “Kate is hugging Beth”) and conjoined sentences (e.g., “Kate and Beth are hugging”). If clear distinctions of grammatical and thematic role modulate assumptions of agency, we might expect to find an effect of first mention on transitive trials and not on conjoined trials. Alternatively, if we find an effect of first mention on agency perceptions for both trial types, including conjoined sentences for which there are no such distinctions between the entities, this would indicate that mention order has an influence above and beyond grammatical role for influencing judgements of agency. We consider this in the context of having already found evidence that mention order biases of relative similarity occur on both transitive and conjoined sentences.

Furthermore, we extend our previous research by referring to an important trend found in Chapter 2; the finding that speakers may show both a man-first bias and a Like Me gender bias in their descriptions of interactions. We question what the implications of this are for comprehension, by also investigating how gender modulates effects of first mention. To do so, we tested the extent to which judgments of agency will differ depending on two important factors. First, we considered how the gender of the mentioned individuals would have an effect, using sentences where a woman is mentioned first and a man is mentioned second, and vice versa (e.g., “Kate is hugging
Fred” versus “Fred is hugging Kate”). We predicted that the first mention agency bias would be stronger when a man’s name was mentioned first rather than a woman’s name.

Second, we tested whether agency judgements on these different-gender sentences will differ depending on the gender of the comprehender. We considered two possibilities; first, that there would be no effect of participant gender. If so, both female and male participants would be more likely to show a first mention agency bias when a man was mentioned first rather than a woman. Following the Like Me gender bias uncovered in Chapter 2, we also considered that on these different-gender sentences, participants would rate the first-mentioned person as the more agentive than the second-mentioned person if the first person was the same gender as themselves (for instance, women being more likely to judge the first-mentioned person as agentive in the case of “Kate is hugging Fred” rather than “Fred is hugging Kate”). The current study therefore expands on our previous work, by exploring the link between social group, mention order and agency.

In order to measure if mention order signifies perceived agency, we were careful to balance the possibility that either person in the event could be the event agent, so that the effect of mention order could be partially dissociated from the meaning of the verb. To do so, we used the same symmetrical verbs as in Experiment 2, describing actions that are being executed by both participants simultaneously (Gleitman et al., 1996), e.g., playing [cards], arguing, dancing with. Unlike asymmetrical verbs, such as maintaining, mentioning or showing (Kako, 2006), symmetrical verbs permit us to present sentences where one figure is not the most obvious event agent. This allowed us to present participants with events where first mention was the more reasonable determiner of agency, rather than meanings derived from the verb (e.g., with showing, one person is doing the showing as the event agent and the other person is being showed as the event
patient, whereas with playing cards, both people are executing the same, symmetric action). We could then measure if participants showed a tendency to interpret the first-mentioned person as more agentive (the person doing the action to the other person) than the second-mentioned person according to mention order specifically, rather than agency as implied by the semantics of asymmetrical transitive verbs.

To measure agency, we asked participants to rate the degree to which each participant caused an event to happen and wanted an event to happen. These properties were chosen based on Dowty’s theory of prototypical agents. Out of Dowty’s properties, we determined that causation and volition lend themselves best to symmetrical verbs because they do not rely on the depiction of different movements or states. Importantly, they also allow us to test the effect of syntax, as symmetrical verbs can be easily used in conjoined sentences (e.g., “Kate and Beth are kissing”) as well as in transitive sentences (e.g., “Kate is kissing Beth”). Therefore, we were able to investigate if biased perceptions of agency can occur based off mention order even when there are no distinctions of grammatical or thematic role between the two people in the description. Thus, we measured agency according to causation (e.g., “Kate asked to kiss Beth”) and volition (e.g., “Kate wanted to kiss Beth”).

As in Kako (2006), participants first read sentences with two nouns and a verb, and then provided subjective ratings on properties of those nouns. Similar to Kako, we rephrased Dowty’s properties to put them into layman’s terms, asking participants to rate the extent to which they considered one figure to have caused the event (on causation trials), and the extent to which they considered one figure to have wanted that event to occur (on volition trials). We predicted that participants would be more likely to rate the first-mentioned person as the more likely causational and volitional agent than the second-mentioned person. We also tested the extent to which syntax would
modulate these perceptions of agency. Lastly, we explored how gender modulates biased judgements of first-mentioned agency by showing sentences where one figure mentioned had a woman’s name and the other had a man’s name.

3.2. Method

We preregistered the experiment on osf.io, with the transitive and conjoined conditions registered as separate experiments (https://osf.io/p9d5v/registrations). There, we specified the number of participants we would test of each demographic group, how we would code responses, the main analyses and exclusion criteria.

3.2.1. Participants

We first conducted the active transitive version of the experiment and then the conjoined version, with 64 participants in each version, therefore 128 adults took part overall (mean age = 38.50 [SD = 15.19], range = 18–74). In both versions, half the participants were female, and half were male, all participants reported that their first language was English, that their location was the UK, they reported no language-related disorders, and had a Prolific approval rating of between 95% - 100%. Participants had not taken part in our previous study on the Like Me bias. The experiment was conducted on Prolific.com and participants were paid £1.88 to take part for 15 minutes.

3.2.2. Materials

Participants provided ratings of causation and volition by clicking a button on a Likert scale. In the transitive condition, critical items consisted of 40 transitive sentences utilising ten symmetrical verbs (the same as in Experiment 2), for example *Fred is playing with Ruth*. These ten symmetric predicates each appeared four times:
twice as “same-gender” trials (one woman-first/woman-second sentence, one man-
first/man-second sentence, n=20) and twice as “different-gender” trials (one woman-
first/man-second sentence, one man-first/woman-second sentence, n=20).

For each sentence, participants provided two ratings: one causation rating and
one volition rating. The phrasing of these ratings was adapted from Dowty (1991), with
causational ratings highlighting one figure as the instigator of the event, and volitional
ratings highlighting one figure as wanting the event to occur, e.g., for the example
above, “Fred asked to play with Ruth” (causation); “Fred wanted to play with Ruth”
(volition). Thus, participants provided 80 ratings in total. Ratings were given on a 7-
point Likert scale where on one end, one figure was positioned as the prototypical agent
(e.g., point 1 = “Fred asked to play with Ruth”) and at the other end, the other figure
was positioned as the prototypical agent (e.g., point 7 = “Ruth asked to play with
Fred”). Participants used the scale to rate which of the two descriptions best matched
the provided sentence. So, for the previous example, they would give a rating between 1
and 4 if they believed that Fred asked to play with Ruth followed best from “Fred is
playing with Ruth”, and a rating between 4 and 7 if they believed that Ruth asked to
play with Fred followed best. In the conjoined condition, the design was the same
except participants viewed 40 conjoined sentences instead of transitive sentences before
providing ratings, for example “Fred and Ruth are playing with each other”.

To check participants were paying attention during the study, they also
completed 8 catch trials, in which they saw a sentence with a perspective predicate
(Gleitman et al., 2007), like “Ruth is chasing Fred”, and rated the accuracy of two
alternative descriptions of the event. Unlike the critical trials, the catch trials were not
subjective. For example, the first rating question might show “Ruth is running away
from Fred” at point 1 and “Fred is running away from Ruth” at point 7, in which case ratings from 5 – 7 were counted as correct.

We included a post-study demographics questionnaire, on which we collected participants’ age, the first language they learnt to speak, the language they currently use the most, the country they currently live in, their nationality, their race, their gender (the latter to check that their gender given in the questionnaire matched the gender they signified on their Prolific account) and if they had lived in the UK for the last 5 years.

3.2.3. Procedure

Participants were asked “to imagine some situations”, by viewing a description of an event involving two figures using either a transitive or conjoined sentence. They then providing ratings of causation and volition for the figures in that event on critical trials, and ratings of semantic accuracy on catch trials.

On critical trials, the initial sentence was presented on-screen for 2000 ms and then re-appeared again twice – once with a scale for rating causation (i.e., which figure initiated the interaction), and once with a scale for rating volition (i.e., which figure wanted the interaction to occur); see Figure 1 for an example of both scale types. We randomised whether causation or volition were rated first. Across lists, we counterbalanced which name was mentioned first in each sentence (e.g., “Fred is playing with Ruth” vs. “Ruth is playing with Fred”) and counter-balanced which names were paired together. We also counterbalanced whether the first-mentioned person appeared as the prototypical agent on the left-hand-side of the rating scales or the right-hand-side (Figure 1).
**Figure 1**

*Examples of a different-gender causation trial (top) and a volition trial (bottom)*

*Note.* This example, using the verb “play”, depicts how the rating scale position was counterbalanced. The first-mentioned person (Fred) appears as the prototypical agent on the left-hand-side of the rating scale in non-boldface, and on the right-hand-side of the rating scale in boldface.

On catch trials, the initial sentence was also presented for 2000 ms and reappeared twice with the two ratings of semantic accuracy. The order of these two ratings was also randomised, as was the order of critical items and catch items.
After completing the language task, participants completed the post-study demographics questionnaire. This experiment was programmed using JavaScript and the jsPsych library (de Leeuw, J. R., 2015).

3.3. Analysis

The design of this study was mixed, involving a within-subjects factor in which all participants viewed the same Likert scale across same gender and different-gender trials, and two between-subjects factors. First, the syntax of the sentences: participants viewed different initial phrases depending on whether they had been assigned to the transitive or conjoined condition. This allowed us to test many trials while keeping the experiment duration down, and to avoid confounding effects incurred by exposure to a different syntax. Second, gender identity: whether the participant’s gender was the same as the first-mentioned person on different-gender trials. For analysis, the 7-point scale was recoded to be -3 to 3, so that 0 was the midpoint representing no difference in agency or volition.

For the main analysis, we measured if participants rated the first-mentioned person as more agentive than the second-mentioned person, looking at combined ratings of causation and volition and each agency property separately. We predicted that participants would rate the first-mentioned person as the causative or volitional agent above chance, indicated by a significant intercept.

The gender analysis focused on different-gender trials. We predicted that participants would be more likely to rate the first-mentioned person as agentive if this person was a man, with a woman mentioned second, rather than a woman, with a man mentioned second. Additionally, we analysed the effect of participant gender (if female
participants and male participants differed in their perceptions of agency) and the interaction between the gender of the first-mentioned person and participant gender, to determine if a Like Me gender bias was present. If so, this would manifest as male participants giving higher agency ratings when the first-mentioned person was a man, and female participants giving higher agency ratings when the first-mentioned person was a woman. Since three-way interactions made the combined model too complex to converge, we restricted this analysis to including two-way interactions only. The separate causation and volition models did converge with three-way interactions, so we were able to test if a Like Me gender bias, if present, was also more likely to occur with transitive sentences versus conjoined sentences.

For both analyses, we also measured the extent to which syntax affected perceptions of agency, comparing transitive and conjoined sentences to determine if agency ratings differed across the sentence types (a main effect of syntax). We predicted that the first-mentioned agency bias would be stronger for transitive sentences than conjoined sentences. We also measured the extent to which perceptions of agency differed in regard to the agency property being rated, comparing causation and volition ratings (a main effect of agency property), and without predicting which property would show the strongest effect.

We also looked at the interaction between these variables (syntax and agency property) for both combined models, allowing us to explore whether differences in how the first-mentioned agency bias manifested for each agency property would be more prominent for transitive sentences rather than conjoined sentences.

All analyses were performed using mixed effects logistic regressions in R (R Core Team, 2019) via the lmerTest package (Kuznetsova et al., 2017). Models included by-item and by-subject random intercepts and the maximal random effects structure that
permitted model convergence. If the models had a singular fit or did not converge, we
removed random effects that accounted for the least variance until the models were
appropriate. Only the final models are reported in the results. We describe each
regression analysis using the lme4 model syntax.

3.3.1. Exclusions

We pre-registered a set of exclusion criteria to ensure data quality. We excluded
one participant for not reporting their gender and seven participants for answering “No”
to having lived in the UK for the last 5 years. We excluded trials where the response
time was 3 standard deviations greater than the mean (2% of trials). We also excluded
participants who got more than 2 out of the 16 catch questions incorrect (n=9). Finally,
we pre-registered to exclude participants who routinely selected the same response on
80% of trials (more than 76 trials). However, this final exclusion criterion was in
retrospect an error– while it only excluded 2 participants in the transitive condition, it
excluded 45 (out of 64) participants in the conjoined subject condition, for which
participants were much more likely to choose the midpoint. All 45 participants selected
the midpoint on more than 76 trials. In fact, the highest number of repetitions for the
mid-point was 80 out of 96 – only a few trials above our threshold. Since that exclusion
criterion could lead to biases in the results, we removed it for the presented analysis.
After exclusions, the total number of participants in this experiment was 112 (female =
55; mean age = 38.05 [SD = 14.57], range = 18–72; 58 in the transitive condition, 54 in
the conjoined condition).
3.4. **Results**

We analysed if participants showed a bias toward perceiving the first-mentioned person as more semantically agentive than the second-mentioned person, by taking ratings on two different features of prototypical agency: causation and volition.

3.4.1. **Primary analysis**

Firstly, we analysed if participants perceived the first-mentioned person as more agentive than the second-mentioned person, and the extent to which this perception was influenced by syntax (transitive/conjoined), agency property (causation/volition) and their interaction. These effects are depicted in Figure 2.
Figure 2

The first-mentioned agency bias

Note. The mean causational and volitional agency ratings in favour of the first-mentioned person, across transitive and conjoined trials. Above zero represents ratings in favour of the first-mentioned person as more agentive, below zero represents ratings in favour of the second-mentioned person as agentive, and zero represents chance. After Likert scale re-coding, rating limits were from -3 to 3. Error bars represent standard error.

We conducted three linear mixed effects models. The first model combined ratings from both causation and volition trials and was of the form Agency Rating ~ 1 + Syntax * Agency Property + (1 | Subject) + (1 | Verb), the second took a subset of the data using causation trials only, and the third took a subset using volition trials only.
(and therefore rating type was not a fixed effect in these latter models). Results are displayed in Table 1.

Table 1

Results from the first-mentioned logistic regressions

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.43</td>
<td>0.08</td>
<td>46.31</td>
<td>5.53</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.21</td>
<td>0.06</td>
<td>106.00</td>
<td>3.50</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Agency property</td>
<td>0.17</td>
<td>0.06</td>
<td>105.85</td>
<td>2.69</td>
<td>0.01 ***</td>
</tr>
<tr>
<td>Syntax : Agency property</td>
<td>0.25</td>
<td>0.06</td>
<td>105.85</td>
<td>3.97</td>
<td>0.00 ***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Causation-only</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.43</td>
<td>0.07</td>
<td>71.64</td>
<td>6.12</td>
<td>0 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.21</td>
<td>0.06</td>
<td>106.00</td>
<td>3.50</td>
<td>0 ***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Volition-only</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>df</td>
<td>t value</td>
<td>p</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.60</td>
<td>0.09</td>
<td>31.54</td>
<td>6.52</td>
<td>0 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.46</td>
<td>0.07</td>
<td>105.98</td>
<td>7.07</td>
<td>0 ***</td>
</tr>
</tbody>
</table>

***p < .001.

We found a significant intercept for all three models. This indicated that, overall, participants rated the first-mentioned person as more agentive than the second-
mentioned person (M = 0.52 [SD = 0.65]), and this held for measurements of causation (M = 0.43 [SD = 0.65]) and for volition (M = 0.60 [SD = 0.81]). These results therefore support our hypothesis that perceptions of agency are influenced by mention order.

We also found a significant main effect of syntax on perceived agency for all three models. Participants were more likely to rate the first-mentioned person as agentive when the initial phrase was transitive rather than conjoined (M = 0.84 [SD = 0.56] vs. M = 0.17 [SD = 0.56]). Mention order had an effect with both syntax types, which we found by analysing the transitive and conjoined sentences separately, using models which were both of the form Agency Rating ~ 1 + Agency Property + (1 + Agency Property | Participant) + (1 | Verb). The first-mentioned person was not only rated as more agentive on transitive sentences ($\beta = 0.64$, $SE = 0.11$, $df = 27.51$, $t = 5.59$, $p < .001$) but also on conjoined sentences where both people had the same grammatical role ($\beta = 0.21$, $SE = 0.09$, $df = 53.50$, $t = 2.33$, $p < .05$).

From the main model, we also found a significant main effect of agency property; ratings for volition were significantly higher than those for causation. This effect was qualified by the significant interaction between syntax and rating type. On transitive trials, participants showed a greater likelihood to rate the first-mentioned person as more agentive than the second-mentioned person for the property of volition rather than causation. For conjoined trials, however, there was little difference between the agency property ratings. This was further identified with the independent models that analysed transitive and conjoined trials separately. On transitive trials, participants were more likely to rate the first-mentioned person as more agentive than the second on volition trials than on causative trials ($\beta = 0.41$, $SE = 0.07$, $df = 57.01$, $t = 5.72$, $p < .001$). However, on conjoined trials, there was no statistical difference between causation and volition ratings ($\beta = -0.08$, $SE = 0.10$, $df = 52.94$, $t = -0.77$, $p = .44$).
3.4.2. Gender analysis

Next, we analysed the influence of gender on perceptions of agency using only the different-gender trials. In this analysis, we investigated effects of first-mentioned gender (if the gender of the first-mentioned person affected the likelihood that participants would perceive that name as being more semantically agentive than the second-mentioned person) and effects of syntax. Figures 3 displays these effects. Additionally, we explored the effect of participant gender and the interaction between these two variables to investigate a potential Like Me gender effect. These effects are displayed in Figure 4.
The gendered first-mentioned agency bias modulated by syntactic form

Note. The mean causational and volitional agency ratings on different-gender trials in favour of the first-mentioned person, across transitive and conjoined trials and trials where the first-mentioned person was a man or a woman. Above zero represents ratings in favour of the first-mentioned person as more agentive, below zero represents ratings in favour of the second-mentioned person as agentive, and zero represents chance. After Likert scale re-coding, rating limits were from -3 to 3. Error bars represent standard error.
The gendered first-mentioned agency bias modulated by participant gender

Note. The mean causational and volitional agency ratings on different-gender trials in favour of the first-mentioned person, across participant gender and trials where the first-mentioned person was a man or a woman. Above zero represents ratings in favour of the first-mentioned person as more agentive, below zero represents ratings in favour of the second-mentioned person as agentive, and zero represents chance. After Likert scale re-coding, rating limits were from -3 to 3. Error bars represent standard error.

As with the previous analysis, we conducted three linear mixed effects models for the combined causation and volition trials, and for causation and volition trials separately (see Table 2 for results). The first model, using ratings from both causation and volition trials, was of the form $\text{Agency Rating} \sim (\text{Syntax} + \text{Agency Property} + \text{First-}$
mentioned Gender + Participant Gender)^2 + (1 + Agency Property + First-mentioned Gender | Subject) + (1 | Verb). We found a significant intercept, indicating that participants were rating the first-mentioned person as more agentive than the second-mentioned person (M = 0.51 [SD = 0.66]). Again, this was the case for both causative ratings (M = 0.42 [SD = 0.66]) and volitional ratings (M = 0.61 [SD = 0.82]).

A significant main effect of first mention gender suggested that participants were also more likely to rate the first-mentioned person as more agentive than the second if this person was a man rather than a woman (M = 0.56 [SD = 0.70] vs. M = 0.47 [SD = 0.65]). Importantly, there was no interaction between first-mentioned gender and participant gender; both male and female participants rated the first-mentioned person as more agentive if they were a man rather than a woman (see Figure 4).
Table 2

Results from first mention logistic regressions on different-gender trials

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.41</td>
<td>0.08</td>
<td>42.16</td>
<td>5.34</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.21</td>
<td>0.06</td>
<td>108.43</td>
<td>3.51</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Agency property</td>
<td>0.18</td>
<td>0.06</td>
<td>108.93</td>
<td>2.90</td>
<td>0.00 **</td>
</tr>
<tr>
<td>First-mentioned gender</td>
<td>-0.07</td>
<td>0.02</td>
<td>317.46</td>
<td>-3.70</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.04</td>
<td>0.06</td>
<td>108.47</td>
<td>0.62</td>
<td>0.54</td>
</tr>
<tr>
<td>Syntax : Agency property</td>
<td>0.25</td>
<td>0.06</td>
<td>108.93</td>
<td>4.09</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax : First-mentioned gender</td>
<td>-0.01</td>
<td>0.01</td>
<td>110.09</td>
<td>-0.97</td>
<td>0.34</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>-0.05</td>
<td>0.05</td>
<td>107.97</td>
<td>-1.00</td>
<td>0.32</td>
</tr>
<tr>
<td>Agency property : First-mentioned gender</td>
<td>0.05</td>
<td>0.02</td>
<td>4052.91</td>
<td>1.92</td>
<td>0.06*</td>
</tr>
<tr>
<td>Agency property : Participant gender</td>
<td>0.12</td>
<td>0.06</td>
<td>108.92</td>
<td>1.95</td>
<td>0.05*</td>
</tr>
<tr>
<td>First-mentioned gender : Participant gender</td>
<td>-0.01</td>
<td>0.01</td>
<td>110.07</td>
<td>-0.52</td>
<td>0.61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.41</td>
<td>0.07</td>
<td>53.70</td>
<td>5.75</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.21</td>
<td>0.06</td>
<td>108.02</td>
<td>3.51</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>First-mentioned gender</td>
<td>-0.07</td>
<td>0.02</td>
<td>108.92</td>
<td>-3.46</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.04</td>
<td>0.06</td>
<td>108.02</td>
<td>0.62</td>
<td>0.53</td>
</tr>
<tr>
<td>Syntax : First-mentioned gender</td>
<td>0.01</td>
<td>0.02</td>
<td>108.92</td>
<td>0.42</td>
<td>0.68</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>-0.05</td>
<td>0.06</td>
<td>108.92</td>
<td>-0.85</td>
<td>0.40</td>
</tr>
<tr>
<td>First-mentioned gender : Participant gender</td>
<td>0.00</td>
<td>0.02</td>
<td>108.92</td>
<td>0.17</td>
<td>0.87</td>
</tr>
<tr>
<td>Syntax : First-mentioned gender : Participant gender</td>
<td>0.01</td>
<td>0.02</td>
<td>108.92</td>
<td>0.38</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.59</td>
<td>0.09</td>
<td>30.64</td>
<td>6.77</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.46</td>
<td>0.06</td>
<td>107.99</td>
<td>7.29</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>First-mentioned gender</td>
<td>-0.02</td>
<td>0.02</td>
<td>2069.34</td>
<td>-1.30</td>
<td>0.19</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.16</td>
<td>0.06</td>
<td>107.99</td>
<td>2.47</td>
<td>0.01*</td>
</tr>
<tr>
<td>Syntax : First-mentioned gender</td>
<td>-0.04</td>
<td>0.02</td>
<td>2069.33</td>
<td>-2.04</td>
<td>0.04*</td>
</tr>
<tr>
<td>Syntax : Participant gender</td>
<td>-0.05</td>
<td>0.06</td>
<td>107.99</td>
<td>-0.77</td>
<td>0.44</td>
</tr>
<tr>
<td>First-mentioned gender : Participant gender</td>
<td>-0.02</td>
<td>0.02</td>
<td>2069.35</td>
<td>-0.97</td>
<td>0.33</td>
</tr>
<tr>
<td>Syntax : First-mentioned gender : Participant gender</td>
<td>-0.02</td>
<td>0.02</td>
<td>2069.35</td>
<td>-1.28</td>
<td>0.20</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .01. • = marginally significant.
This analysis also recapitulated our findings of a significant main effect of syntax in the combined model. This perception bias was stronger on transitive trials than on conjoined trials (M = 0.84 [SD = 0.58] vs. M = 0.17 [SD = 0.57]). Agency property was also a significant predictor, with participants giving the first-mentioned person higher ratings of volition than causation (M = 0.61 [SD = 0.82] vs. M = 0.42 [SD = 0.66]).

Again, the interaction between syntax and agency property was significant, in the same direction as in the first analysis. Volitional ratings of the first-mentioned person were higher than causational ratings in the transitive condition, while there was little difference between the agency properties in the conjoined condition. Interactions between agency property and first-mentioned gender, and agency property and participant gender were marginally significant. No other interactions were significant.

We ran two further models, whose results are also displayed in Table 2 and trends are also depicted in Figure 3. These analysed either causation trials only, or volition trials only. The causation model was of the form Rating ~ Syntax * First-mentioned Gender * Participant Gender + (1 + First Mentioned Gender | Subject) + (1 | Verb). The volition model omitted the random effect of first mentioned gender for model convergence. Results from both models showed a significant intercept. Participants reliably rated the first-mentioned person as the more causative or volitional agent respectively than the second-mentioned person. Results also showed a significant main effect of syntax for both models; these causative and volitional biases were stronger on transitive trials than on conjoined trials (causative: M = 0.63 [SD = 0.61] vs. M = 0.20 [SD = 0.66]; volitional: M = 1.05 [SD = 0.66] vs. M = 0.13 [SD = 0.70]).

In the causation model, we also found a significant main effect of first-mentioned gender. Participants were more likely to rate the first-mentioned person as
the more likely causative agent than the second-mentioned person if the first-mentioned person was a man rather than a woman (M = 0.49 [SD = 0.73] vs. M = 0.35 [SD = 0.65]). We did not find any significant interactions in the causation model. In the volition model, and notably, we found a main effect of participant gender. Female participants were significantly more likely to rate the first-mentioned person as the volitional agent than male participants (M = 0.75 [SD = 0.89] vs. M = 0.46 [SD = 0.72]). Participant gender had not been a significant predictor in previous models.

Unlike in the combined and causation model, first-mentioned gender was not a significant predictor of volitional agency perceptions. Participants were not generally more likely to rate the first-mentioned person as the volitional agent if this person was a man rather than if this person was a woman. However, the volition model showed a significant interaction between syntax and first-mentioned gender. When the gender of the first-mentioned person did influence agency judgements, this occurred on transitive trials, with first-mentioned person receiving higher volitional ratings when male than female (M = 1.10 [SD = 0.66] vs. M = 0.99 [SD = 0.71]). On conjoined trials, the difference was negligible between man-first orders and woman-first orders (M = 0.12 [SD = 0.80] vs. M = 0.14 [SD = 0.65]).

3.5. Discussion

The results from this experiment provide new evidence on how the linguistic factors of grammar, mention order, and gender influence our judgements of social agency, with participants showing a reliable tendency to rate a person who is mentioned first as a more likely agent than a person who is mentioned second. Our syntax condition revealed crucial implications for the importance of grammatical function and thematic role on these judgements. We found a very large effect in the transitive
condition, as predicted. In these sentences, order of mention is confounded with syntactic and semantic role, so we cannot be certain that it is mention order as such driving these effects. However, by using symmetrical predicates, we were better positioned to test judgements of social agency than if we had used sentences where agency is more implied on account of one noun clearly acting on or reacting to another (as in one of our earliest examples where the hiker ran away from the bear).

However, and crucially, we also found a smaller but nonetheless significant effect on conjoined trials, indicating that distinctions of grammatical and thematic role are not essential for making semantic inferences of agency from mention order. In other words, this finding suggests that mention order has an independent effect on agency perceptions above and beyond grammar; a finding which is consistent with Kesebir’s work on binomials and relevance (Kesebir, 2017). Still, the effect of first mention on agency judgements was larger for transitive sentences than for conjoined sentences, which suggests these interpretations are more likely to occur when grammatical and thematic role distinctions are present in the syntax.

Our data allowed us to break down judgements of agency into two components: perceived causation and perceived volition. In most ways, these two components behaved in lockstep, but there were some interesting and theoretically informative differences. Figure 2 indicated that, on transitive trials, the first-mentioned person received higher ratings of volition than causation. Participants tended to rate the first-mentioned person as having wanted the event to occur more often than they rated the first-mentioned person as having caused the event. On conjoined trials, however, there were no clear differences between these agency property ratings. This not only demonstrates that there are important perceptual differences between perceived...
causation and volition, but that those differences may appear more obvious when there are also clear distinctions of thematic role.

However, it may also suggest a possible weakness in our chosen phrasing of causation ratings. Whereas all volition ratings followed the same format for each verb (Kate wanted to [verb] + [preposition] Beth), the format of causation ratings differed across verbs for the purpose of comprehension and naturalness (e.g., for the verb marry, the causation rating was “Kate asked to marry Beth”, while for the verb flirt, the causation rating was “Kate started flirting with Beth first”). The effect of agency property on agency judgements also highlights the necessity to explore other properties of perceived agency in relation to mention order, not just causation and volition, and to test a range of different symmetrical verbs.

Our previous work on the Like Me bias revealed that we tend to talk about dyadic interactions from the perspective of the person more similar to us, but also that when dyads differ in gender, this is at times modulated by an androcentric “man-first” bias that shows both men and women may mention a man before a woman. To follow this up, we tested the extent to which gender modulates the effect of first mention on agency judgements in the current study. As predicted, on different-gender trials participants were more likely to rate the first-mentioned person as the more likely agent when this person was a man rather than a woman, suggesting that androcentrism also exerts a role in the comprehension of event descriptions. Consistent with Hegarty and colleagues (2011), we did not find an effect of participant gender on agency judgements for different-gender sentences. Both female and male participants showed a greater likelihood to rate the first-mentioned person as more agentive than the second if the first-mentioned person was a man and the second-mentioned person was a woman (e.g., Fred is playing with Kate). Thus, the Like Me bias that we have documented in
production is not seen clearly in comprehension; instead, we see evidence of a male-dominant world view.

These findings do highlight an important implication of the Like Me bias, however, for the way in which information is transmitted. Depending on who is speaking (or writing), there may be biases in order of mention and perspectives taken, which can then have effects on judgements of roles. What does this mean for areas where certain voices are overrepresented? One example of this can be found in literature. Looking specifically at children’s literature and writing, Hsiao and colleagues drew links between the overrepresentation of male authors writing children’s books with children showing a bias towards writing stories about boys. They also found that this bias increases with age (Hsiao et al., 2021). Furthermore, Johns and Dye (2019) found that authors show an overwhelming bias to use names for men in their writing over names for women, in both non-fiction and fiction, and across a range of historic and current written media. If there are more male protagonists than female protagonists, what does this say about how reading might influence gendered perceptions of agency? A meta-analysis of literature in the third-person narrative is beyond the scope of this paper, however this may be an interesting area for further study.

The gender effects of the first-mentioned agency bias have considerable importance because it is often the case that men are mentioned first and women second. For instance, Hegarty and colleagues (2011) have found that when people name unfamiliar heterosexual couples, they tend to mention the man’s name first. In a discourse analysis of a 5th grade English Language textbook, Ahmad and Shah (2019) identified that, in instances where women and men were mentioned together in the same sentence, the man was mentioned first 65% of the time. They argue that the way in which gender is represented in educational texts for children may assert a male-biased,
androcentric world view and contributes to the fixed constructs of gender during development and into adulthood. Certainly, our results suggest that when a man-first bias is present, a reader is more likely to perceive the man as more agentive than the woman, giving further justification to Ahmad and Shad’s concerns.

Why does first mention inspire judgements of agency? One explanation is that, for transitive sentences and in languages where the Subject-Verb-Object structure is common, people learn to link sentence subjects with agents through linguistic experience (Foley & Van Valin, 2009; Bowerman, 1990). Thus, someone might infer that the first-mentioned noun in a transitive sentence is the semantic agent, because their statistical assumptions encourage this expectation. We might also consider that these judgements are a matter of pragmatics; that a comprehender might assume a speaker would mention the most important element first (Grice, 1975).

The effect of first mention on agency judgements has important societal implications. Re-ordering commonly fixed orders, such as when men are mentioned first and women second, could help restore some balance for underrepresented or stereotyped groups. However, just as it could be used to amplify agency in a productive way, first mention has the potential to skew perceptions of agency with destructive consequences, including creating perceptions of blame, culpability and responsibility. For example, research that has analysed the way in which victims are spoken about has revealed that mentioning the victim first in passive sentences has the potential to shift the blame from the perpetrator to the victim (Fast & Kinewesquao, 2019). In the case of sexual violence, the positioning of the victim at the forefront of a passive sentence influences perceptions of the victim’s responsibility; a perception that is even more likely from those who are already more accepting of “rape myths” (Bohner, 2001).
Similar patterns could be attributed to other forms of violence. For example, the stimuli and design of this experiment did not make it appropriate for investigating the effects of race on agency perceptions, but we have already seen the importance of race as well as gender with respect to first mention in our Like Me production experiments. We can therefore consider how these patterns might relate to the constantly pressing cross-continental issue of police violence against Black people.

To illustrate this point, in 2020 the UK Prime Minister at the time, Boris Johnson, said the following in a speech after the death of George Floyd – a Black man whose death while in police custody that same year inspired protests across the world.

“The death of George Floyd took place thousands of miles away – in another country, under another jurisdiction – and yet we simply cannot ignore the depth of emotion that has been triggered by that spectacle, of a black man losing his life at the hands of the police”

By framing the violence in this way, mentioning “a black man” first as the sentence subject and the police second, focus is shifted onto the victim and away from the police, and thus the societal problem of police brutality and racism. However, there may also be times where the opposite, for example “of the police killing a black man”, may be inappropriate, for instance where the intention may be to focus on the harm that came to George Floyd and to reflect on his personhood. Thus, mention order must be addressed according to the complexities that it can hold; that first mention can be both constructive and damaging, and that highlighting agency can be both desirable and undesirable. Future studies could use the paradigm provided in this experiment to investigate how speakers and writers use the passive voice to shift subject focus as it relates to perceptions of agency, as well as exploring the effect of race on first mention more generally.
This study highlights the importance of not taking mention order at face-value – that hidden meanings may exist in even the simplest of event descriptions and that the biases we hold as a society are likely influencing how we make assumptions from the perspectives we are given.
4. Study 3: Animacy and human-robot interactions\textsuperscript{10}

4.1. Introduction

The animacy bias is a widely tested and accepted phenomenon of language production. When choosing how to order nouns in a sentence, speakers tend to mention the most animate entity first and the less animate entity second (Branigan et al., 2008), for example, “the hiker ran away from the bear”, even when this requires using a more complex syntactic structure such as a passive sentence. This bias has been identified in both adults and children (Lempert, 1989; Buckle, 2020; Buckle et al., 2017), suggesting that animacy becomes an important determiner of production decisions at an early developmental stage.

The animacy bias is argued to occur through its association with conceptual accessibility, signifying how easily a concept can be retrieved from memory (Bock, 1986a; Bock & Warren, 1985; McDonald et al., 1993). An entity’s conceptual accessibility is determined by how many relationships it can enter into or, predicates that can be ascribed to it, which ultimately correlates to how many things it can do. Since animate entities are able to do more than inanimate entities, they are more conceptually accessible. The relative conceptual accessibility of animates and non-animates corresponds closely to the animacy hierarchy proposed by Harris (1978) and Aissen (2003), with abstract concepts at the bottom of the hierarchy, followed by non-animate and non-living objects (e.g., rocks), then non-animate living objects (e.g., plants), followed by non-human animate objects (e.g., animals) and lastly humans at the top. When predicting how words will be ordered in sentences, the hierarchical framing

\textsuperscript{10} Experiments 6 and 7 in this study were designed by the author in collaboration with Dr. Anna Henschel and Prof. Emily S. Cross at the University of Glasgow.
of animacy provides a tidy prediction for mention order. Entities higher up on the animacy, or conceptual accessibility, hierarchy tend to be mentioned earlier than entities lower on this hierarchy (Branigan et al., 2008; Prat-Sala & Branigan, 2000). Thus, relative animacy tends to determine which entity a speaker will mention first.

But what makes an object animate, and is this quality as fixed as the hierarchy might imply? In order to fully explain why a speaker is more likely to say “the hiker ran away from the bear” rather than “the bear chased the hiker”, we must explore the properties that distinguish hiker from bear. In this study, we show that the way in which people conceptualise animacy is extremely complex and multi-faceted. Importantly, we also demonstrate that it is not as fixed as previous work has suggested, and that it can instead be engineered through information provided in a narrative.

Crucially, the property of animacy is tied to the degree to which we consider or make assumptions about an object’s mental state and therefore its ability to do certain things (e.g., considering that a plant has little or no mental state while a person does). Independent, self-initiated motion is considered a key element in determining animacy and mental states, as it may be a marker of intention (Heider & Simmel, 1944; Blakemore & Decety, 2001; Scholl & Tremoulet, 2000; Tremoulet & Feldman, 2006). People make inferences about the internal mechanism responsible for such movement (Schultz & Bülthoff, 2019), whether that is a mind, or a set of rules programmed in a machine. Intention is therefore an important cognitive feature of how social interactions are perceived, relating to both the representation of a figure’s covert mental state and the judgments someone makes from perceiving an action a figure is carrying out (Castelli et al., 2000). In this way, intentionality also relates to the concept of agency, representing properties such as causation (causing another object to experience an event or change in state), volition (choosing to be involved in the event or state change) and
independent movement (Dowty, 1991). For example, a person may create a set of assumptions on witnessing a hiker running away from a bear, such as considering that the hiker may be avoiding being eaten.

Closeness to humanness is another way in which animacy is conceptualised, as the category of “human” sits at the top of the animacy hierarchy (Aissen, 2003; Harris, 1978). This closeness could refer to how human-like a non-human animate entity is. In psycholinguistics, this comes down to the number and types of relations the entity can enter into (Branigan et al., 2008), such as a dog that can respond to commands versus a centipede that cannot. Consequently, a dog is more conceptually accessible than a centipede.

Closeness to humanness could also refer to the way in which non-human or even non-animate entities such as objects can be anthropomorphised, meaning attributing human-like qualities to them such as physical form, abilities and actions, and other characteristics (Bartneck et al., 2009). For instance, we see this in the way eyes have been added to a Henry Hoover in order to create an appealing or even cute household appliance; hoovers without faces would be comparatively less anthropomorphic. Likewise, Noo-Noo, the vacuum cleaner who moonlights as a housekeeper and guardian in the children’s television show Teletubbies, is more anthropomorphic than a Henry Hoover, because it exhibits some degree of facial expression, makes human-like slurping noises, and interacts with the Teletubbies, as well as performing the all-important functions of a normal vacuum cleaner. Thus, it can enter into more relations than a Henry Hoover, meaning it has a higher predicability, and so we could say that Noo-Noo is more conceptually accessible than a Henry Hoover.

Human animacy, that which distinguishes humans from other animals, is constructed from physical attributes, such as form and movement, and a set of beliefs
about mental states and abilities (Cross et al., 2016). Importantly, it relates to things that humans can do that no other species in the animal kingdom is capable of, such as advanced planning or the expression of complicated emotions. Thus, the more human-like an entity’s abilities and appearance are, the closer to humanness it will be perceived.

This paper seeks to explore the effect of animacy on language production by studying descriptions of interactions between humans and robots. The reason behind involving robots in this investigation was two-fold. Firstly, robots represent an interesting tier on the conceptual accessibility hierarchy. On account of not being alive, we might consider that they are less animate than non-human animals. However, robots, or social robots to be specific, are often engineered to have anthropomorphic, human-like features, for example, those that have faces and limbs, those that can move in complex, unpredictable manners, and those with voices (Bartneck et al., 2009; Broadbent, 2017; Cross et al., 2016). This proximity to humanness, without the biological elements that characterise other animals as animate, make robots an interesting case study in the topic of language production that is animacy-driven. Furthermore, the way in which they can be anthropomorphised and given human-like behaviours allows us to test the claim that predicability is what drives mention order biases, rather than animacy per se.

Secondly, robots are becoming increasingly more common in society, from the seemingly simplistic household hoovers or the ‘Roomba’ to the more complex stand-ins for waiters at restaurants and robots who have been introduced to civil forces such as the police. How language is used with or about robots is still relatively unclear – there is so far not much crossover between social robotics and psycholinguistics, but there is a
strong likelihood that people will be interacting with robots much more in the future, making this an important area of study.

In the domain of social cognition, Gray et al. (2007) provide a useful way to categorise differences between entities, by identifying two specific dimensions of mind perception; agency and experience. Here, agency is defined as an object’s ability to plan and act, while experience refers to an object’s ability to sense and feel. In psycholinguistics mind abilities could be understood as contributing to an entity’s inherent accessibility. This is a factor of conceptual accessibility that is determined by semantic characteristics such as animacy and prototypicality (Prat-Sala & Branigan, 2000). Thus, entities with higher animacy and fewer mind abilities will be more inherently accessible than entities with lower animacy and fewer mind abilities. Gray et al.’s definitions of agency and experience are closely linked to the notions of agent and experiencer in linguistics, which differentiate between the ways in which verbs are categorised according to thematic roles, and therefore predicates, that they are associated with. For instance, agents are associated with agent-patient verbs (e.g., to push) where the agent appears as the subject in an active sentence, whereas experiencers are associated with experiencer-theme verbs (e.g., to fear) and theme-experiencer verbs (e.g., to please). A thematic agent might act on a patient, while an experiencer undergoes a change in mental state through having a given experience (Dowty, 1991; Stevenson et al., 1994).

In a study measuring judgements of mind perception, Gray and colleagues collected ratings of a variety of characters, including people and robots, identifying where each one lay on a crossed scale from no experience to high experience, and from no agency to high agency. People were generally rated as having both high agency and experience, whereas robots were rated as having moderate agency but no experience.
(Gray et al., 2007). Henschel et al. (2021) provide additional evidence for these differences of mind perception between humans and robots, again by collecting ratings of these dimensions for humans, robots and other objects. As in Gray et al., participants rated humans as high for both dimensions and ascribed robots with some agency but no experience (while objects had the lowest ratings for both agency and experience). Thus, the difference in how people perceive humans and robots is highly related to the capabilities people ascribe to either group, for instance to experience emotions, to be intentional in their actions, and to maintain beliefs and express aspirations, as these are all central to what makes something more or less human.

Neuroimaging studies provide some understanding as to how humans and robots are processed differently. For example, Wang and Quadflieg (2015) compared participants’ responses to interactions between either two humans or between a human and a robot. They identified that different regions of a neural network involved in formulating impressions relating to beliefs, desires, motives, feelings, or intentions (also known as the social brain network) were activated depending on whether the interaction was human/human or human/robot. Their results show that perceiving the actions of robots, even those that are human-like, stimulate different cognitive responses to those that arise from perceiving the actions of people, indicating that this network is reserved for perceiving other humans.

Empathy, the ability to share and understand the subjective experiences and feelings of another person (Wispé, 1986), is an important feature of mentalising, or the ability to recognise and understand the mental states of others, and of seeing the world from another’s perspective (Frith & Frith, 2006). In order to empathise with someone, or something, you must acknowledge that they have the capacity to feel. Studies have found that, given the right circumstances, we are capable of demonstrating empathy for
robots as well as humans (Küster & Swiderska, 2021; Rosenthal-von der Pütten et al., 2013). Riek and colleagues (2009) have provided evidence for the effectiveness of anthropomorphism for showing empathy towards robots. When shown examples of humans mistreating robots, such as shouting at them, pushing them, or ordering them around unkindly, the likelihood that participants would show empathy towards the robots increased the more “humanlike” their qualities were (e.g., an adult-sized android over Roomba the robot vacuum cleaner). Additionally, Gonsior et al. (2011) discovered that empathy towards a robot during human-robot interactions was highly influenced by the robot carrying out the social behaviour of facial-expression mirroring with its human partner, thereby increasing its perceived humanness.

The current study asks what differentiates humans from non-humans, testing the role of human animacy on mention order. In Experiment 6, we firstly determined if mention order can be predicted by relative animacy between humans and robots using a picture description task. Participants viewed events of human and robot dyads interacting and described these interactions using transitive sentences. We predicted that participants would demonstrate a robust human-first bias in their descriptions, mentioning the human first and the robot second.

Participants were prompted to use symmetrical predicates in their descriptions. As discussed in Chapters 2 and 3, symmetrical predicates are a useful tool because they allow for the possibility of describing an interaction from the perspective of either figure (Gleitman et al., 1996), for example hugging, arguing and playing (the latter relating to playing a card game, rather than an instrument). Thus, decisions of who to mention first as the sentence subject should result from decisions of the most appropriate agent, rather than other factors such as the linguistic accessibility of describing the interaction with a more common verb over a less common verb (e.g.,
possible imbalances between running away and chasing in the hiker/bear example), or more common syntax over a less common syntax (e.g., the transitive is running away from versus the passive is being chased by).

We sought to test the animacy bias further by varying the anthropomorphism of the robot figures, thereby testing gradients of animacy. Anthropomorphism is another means by which an entity could be more, or less, conceptually accessible, as being closer to humanness may imply that it is capable of doing more things than an entity that is further away from humanness. Thus, anthropomorphism establishes an entity’s inherent accessibility. Trials were split between a human paired with a humanoid robot, deemed as the more anthropomorphic robot type on account of having more human-like features, and a human paired with a Lego robot, deemed as less anthropomorphic on account of having fewer human-like features. On human/Lego trials, the gap in animacy between the two figures is larger than on human/humanoid trials. We refer to this as an ‘animacy disparity’; there will be a smaller animacy disparity between a human and a highly anthropomorphic robot and a human, compared to a human and a less anthropomorphic robot. We predicted that participants would mention the human first more often on trials where there is a larger animacy disparity, i.e., when the human is paired with a Lego robot rather than with a humanoid robot.

We formalised the condition of anthropomorphism during a norming study, conducted during a pre-test before the first language experiment. In this study, we collected ratings of properties of the robots that would be depicted as interacting with humans in the language experiments. Participants rated four robots, which were two humanoid robots and two Lego robots, for characteristics of agency and experience, using definitions from Gray et al. (2007). Participants gave similar ratings of both agency and experience to the two humanoid robots, which were higher than the similar
ratings given to the two Lego robots. Thus, we used these ratings to categorise the humanoid robots as more anthropomorphic and the Lego robots as less anthropomorphic.

In the same norming study, participants also provided ratings of gender characteristics for a selection of monosyllabic robot-related fictional names (e.g., Botz, Cybe), so that we could select the most gender-neutral names for the four robot figures. This was done to avoid potential effects of gender, as found in our previous work in Chapters 2 and 3, on first mention. As in Experiments 1 – 5, human figures were assigned monosyllabic, common English names (e.g., Fred, Kate).

Our previous production experiments have indicated that mention order biases may be modulated somewhat by the types of dyadic interactions the speaker is describing. The events depicted in this experiment (and in Experiment 7) were the same actions as depicted in Experiments 1 – 4. Therefore, we used ten symmetrical verbs, which we categorised into three valences: intimate, casual and negative. Different perceptions of animacy are elicited depending on the event (Salatiello et al., 2021), and it is possible that a speaker would respond differently to a human engaging in an aggressive interaction with a robot such as fighting, compared a human engaging in a romantic interaction such as kissing, or a more casual interaction like talking. Thus, we investigated the extent to which the type of event described would influence a human-first bias.

The presence of robots in domestic or work life and in media differs generationally, increasing in prominence as technology has advanced. Thus, exposure to robots could be thought to be influenced by generation, with younger people having grown up with more examples of human-robot interactions. Thus, younger generations may possibly find it more reasonable to view a robot as having agency or experience,
for instance, than older generations, however this has not yet been thoroughly tested. In addition to exposure, it is possible that age may also affect attitudes towards robots. This has been found in certain literature, with older people showing more negative attitudes towards robots than younger people (e.g., Scopelliti et al., 2005), while other research has found no difference in attitudes across these age groups (e.g., Destephe et al., 2015; Kuo et al., 2009). Therefore, the way in which age may affect mention order in this task is unclear, and so we seek to compare older and younger participants for their mention order choices.

Attitudes towards robots is an interesting element of social robotics to consider in relation to mention order choices, as attitudes towards social agents more generally has not yet been studied extensively as a predictor of first mention. We might consider that holding positive attitudes about someone would make a speaker more likely to mention them first. Likewise, holding negative attitudes towards someone may encourage a speaker to describe an interaction from the more favoured person instead. Conversely, overtly negative attitudes may actually increase the salience of a person, therefore influencing a speaker to mention them first. To explore this idea, participants also completed a post-study questionnaire in which we collected measurements of their attitudes towards robots. Items were taken from the Negative Attitude toward Robots Scale (NARS) developed by Nomura and colleagues (2006). This scale addresses three dimensions; Negative attitude toward interaction with robots (which we term ‘general attitude’, 6 items); Negative attitudes toward the social influence of robots (hereby referred to as ‘social influence’, 5 items) and Negative attitude toward emotional interactions with robots (hereby referred to as ‘emotional interactions’, 3 items).

In our previous work on the Like Me bias (Chapter 2), we found that people are more likely to mention a figure who is of their own gender or their own race first, over a
figure who is of a different gender or race. In the norming study, participants also gave ratings of perceived gender of the four robots. As part of an exploratory analysis, the current study also investigated the potential effect of participant gender in relation to perceived gender of the robots, in the context of the Like Me gender bias. We investigated if female participants would be more likely to show a robot-first bias on trials depicting a robot rated as being feminine or female, and if male participants would be more likely to show a robot-first bias on trials depicting a robot rated as masculine or male.

The current study uses the same images of people as in Like Me Experiments 2, 3 and 4, maintaining four demographic groups of humans – Black women, Black men, white women and white men, all paired with a robot. Thus, we conducted an exploratory analysis to determine if a Like Me bias might influence this task, with participants demonstrating a stronger human-first bias on trials where the depicted human was the same gender or the same race as themselves (e.g., a Black participant being more likely to show a human-first bias on trials where a Black person was paired with a robot, than on trials where a white person was paired with a robot). Lastly, control analyses investigated if mention order behaviour changed over the course of the experiment.

In Experiment 7, we sought to measure the degree to which proximity to human-like mind affects mention order. This experiment investigated if it is possible to manipulate beliefs about animacy in order to alter a human and non-human entity’s inherent accessibility. If so, we questioned if presenting a non-human as more animate than a human would consequently reverse the human-first bias. Here, we extended our investigations into the effect of relative social similarity on first mention, which we first uncovered in Chapter 2. We aimed to manipulate relative similarity by presenting the
robots as more socially similar to participants using anthropomorphism and the humans as less socially similar using dehumanisation, thereby changing perceptions of animacy for these groups and their inherent accessibility.

Beliefs are regularly moderated, or suspended, through storytelling. In literature, this is often done through anthropomorphism. Examples include stories about planets that can demonstrate human qualities of resentment and revenge, as in N.K. Jemisin’s The Broken Earth Series, or dogs that can talk and uncover criminal mysteries, as in the beloved Scooby Doo series created by Joe Ruby and Ken Spears. Thus, we used the same paradigm and mostly the same stimuli as in Experiment 6, except we also included a narrative manipulation. Previous research has found that context can affect an entity’s conceptual accessibility, for instance by increasing its salience through discourse, thereby contributing to its short-term derived accessibility (Prat-Sala & Branigan, 2000). Here, we sought to use a short story to alter semantic characteristics of inherent accessibility, with the intention of positioning the robot group as more animate and thus capable of doing more than the human group.

Through this narrative manipulation, Experiment 7 aimed to explore what makes an entity animate. Participants were divided into two conditions: human-oriented and robot-oriented. Participants in a human-oriented condition read a story in which the human figures were given descriptions indicating high levels of agency and experience. The story also aimed to create a sense of empathy for the human figures, who had been given the important task of creating the robot figures via cloning for scientific research and defence.

Participants in a robot-oriented condition read a story that positioned the robot figures as having high levels of agency and experience, aiming to encourage empathy towards the robots, whose task was to create human-like figures via cloning. Thus, the
robot-oriented narrative anthropomorphised the robot group, but also sought to actively dehumanise the human group, describing them not as people, but as clones with limited faculties. Dehumanised perception, which short-circuits empathy (Cameron et al., 2016), results when someone fails to engage typical social cognition and automatically consider the mind of another person (Harris & Fiske, 2011). By presenting the human group as clones in this way, we cued the mechanistic form of dehumanisation, which attributes characteristics of machine-like shallowness, rigidity and lack of agency to people (Haslam, 2006), and we tested if this would make speakers less likely to mention the human first.

Through this narrative manipulation, we aimed to modulate the human-first bias that is expected in Experiment 6, on the grounds that the robot figures in the robot-oriented condition will be more animate and more inherently accessible than the humans. Thus, we predicted that participants in the robot-oriented condition would be less likely to show a human-first bias than participants in the human-oriented condition, and that they might even show a robot-first bias instead.

We tested three different features of animacy for their involvement in conceptual accessibility: agency, experience and empathy. These properties have been linked with animacy, thereby contributing to an entity’s inherent accessibility. Participants provided ratings for each of these animacy properties, for both the human and robot group, in a post-study questionnaire. Perceived agency and experience were measured using items from Gray et al. (2007), including 7 items measuring agency perceptions and 11 items measuring experience perceptions. We measured empathy using the three items from the experimentally validated instrument developed by Seo et al. (2015) to test empathy towards robots and an original item (“How empathetic do you feel towards the [group]?”). Participants also gave ratings of dehumanisation for the human group only,
using items from Harris and Fiske (2011). This was a two sub-scale measure with items addressing general traits (e.g., “How likely are the [story name for human group] to experience the ups and downs of life?”) and story traits (e.g., “I felt like the [story name for human group] in the description were rational and logical, like they were intelligent”). Scores from the post-study questionnaire were used to test the extent to which relative agency, relative experience and relative empathy (between the human and robot groups), and dehumanisation are predictors of mention order.

We anticipated that the degree to which participants perceived the robots as being more animate than the humans could predict the degree to which participants would mention the robot first and the human second. For example, we predicted that participants who empathised more with the robot group than the human group would be more likely to mention the robot first and the human second. Overall, we expected our narrative manipulation to drive these effects, with participants in the robot-oriented condition showing a much weaker robot-first bias than participants in the human-oriented, possibly even a robot-first bias. Thus, we measured if the effects of our narrative manipulation could predict the degree to which participants showed a human or a robot-first bias.

Importantly, we also compared the effectiveness of the narrative manipulation by comparing ratings of agency, experience and empathy of both the human and robot. Participants in the human-oriented condition were expected to give higher mind ratings for the human group than the robot group. If our narrative was successful, we would expect participants in the robot-oriented condition to give higher ratings of agency, experience and empathy for the robot group than the human group. Dehumanisation scores would also be a key determinant of our narrative’s effectiveness, as the human figures are supposed to only look like “real humans” in the robot-oriented condition,
lacking most human qualities other than aesthetics. Thus, we determined that our manipulation was successful if participants in the robot-oriented condition rated the human figures as more dehumanised than participants in the human-oriented condition.

Through Experiment 7, we also sought to replicate the main analyses and findings from Experiment 6, considering the novelty of this work. This time, the figures in both human and robot groups were given fictional names (e.g., Botz, Frez) to avoid effects of linguistic accessibility, as the previously used human names would be much more known to participants than our novel robot names.

4.2. **Experiment 6**

This experiment set out to test if participants would show an animacy bias to mention the human first when describing human-robot dyadic interactions, and the degree to which this was modulated by robot anthropomorphism, event valence and gender.

4.2.1. **Method**

We preregistered the experiment on osf.io (https://osf.io/a5fby/), where we detailed the number of participants we would test, any exclusion criteria, our method for coding responses and the planned analyses.

4.2.1.1. **Participants**

We firstly carried out a norming study for our stimuli on Prolific.com with 76 UK localised English-speaking adults (17 Black women, 17 Black men, 20 white women and 20 white men after exclusions, mean age = 34.16 [SD = 10.73], range = 18 -
68). Demographic groups were chosen on account of matching the same demographic groups as depicted in the human stimuli.

Using Prolific.com, we paid a new group of 240 English-speaking adults localised in the UK for at least the last 5 years £4 to take part in the main experiment for 30 minutes. 60 participants reported that they were Black women, 60 Black men, 60 white women, and 60 white men, mean reported age was 32.77 [SD = 11.92] with a range of 18 – 73. We chose our sample size based on Experiments 1 and 2 as these used a similar paradigm and the same number of participants and trials. Using the procedures described in Westfall, Kenny & Judd (2014) and the online application jakewestfall.shinyapps.io, we determined that we would be able to find a minimum effect size of 0.23 with 80% power. We were therefore sufficiently well-powered to find even a small effect with this number of participants and this number of trials.

4.2.1.2. Pre-test materials

In the norming experiment, participants viewed images of four robots (two humanoid robots, Pepper and NAO by Softbank Robotics, and two Lego Bionicle action figures, ‘Gresh’ and ‘Princess’, Figure 1). They were told these robots were human sized to correct for the real-life differences in size between the Softbank Robotics and Lego Bionicle figures, and between these robots and the people depicted.
Participants then provided ratings of perceived robot agency and perceived robot experience for these four robots, using definitions of agency and experience from Gray and colleagues (2007). They rated the extent to which they considered each robot to have the ability to plan and act, from “No agency” to “Full agency”, and the extent to which they considered the robots to have the ability to sense and feel, from “No experience” to “Full experience”, both using a 9-point Likert scale. Results provided us with mean ratings of perceived agency and experience for each robot.

Figure 2 displays these ratings, showing that the two humanoid robots were rated the highest for agency and experience (NAO agency = 6.08 [SD = 2.03], experience = 5.85 [SD = 2.35]; Pepper agency = 5.91[SD = 2.09], experience = 5.41 [SD = 2.38]), compared to the two Lego robots (Gresh agency = 5.05 [SD = 2.26], experience = 3.69 [2.18]; Princess agency = 4.95 [SD = 2.23], experience = 3.76 [SD =
We explored these differences further with statistical analysis using mixed effects logistic regressions in R (R Core Team, 2019) via the lmerTest package (Kuznetsova et al., 2017). Results revealed that the humanoid robots received significantly higher agency and experience ratings than the Lego robots (agency: $\beta = -0.5$, $SE = 0.09$, $df = 221.00$, $t = -5.23$, $p < .001$; experience $\beta = -0.95$, $SE = 0.10$, $df = 221.00$, $t = -9.69$, $p < .001$).

These graphs also suggest that the humanoid robots could be grouped together based on their similarities of perceived mind, as could the Lego robots. We used these humanoid/Lego groupings as our variable of anthropomorphism for predicting the human-first bias.

Participants also rated each of the four robots for sex and gender characteristics, grading them from very female to very male and very feminine to very masculine on a 9-point Likert scale. Below 5 indicated a more female or feminine rating and above 5 indicated a more male or masculine ratings. Participants showed a tendency to rate the humanoid robot Pepper as female and feminine (mean sex = 2.41 [SD = 1.51], mean gender = 2.45 [SD = 1.38]) and the humanoid robot NAO as slightly male and masculine (sex = 6.12 [SD = 1.67], gender = 5.95 [SD = 1.33]). The Lego robots were rated as very male and masculine in contrast (Gresh sex = 7.09 [SD = 1.84], gender = 7.11 [SD = 1.74]; Princess sex = 8.00 [SD = 1.19], gender = 8.12 [SD = 1.11]). A linear mixed effects model showed that the Lego robots received significantly more male and masculine ratings than the humanoid robots (sex: $\beta = -0.5$, $SE = 0.09$, $df = 221.00$, $t = -5.23$, $p < .001$; experience $\beta = -0.95$, $SE = 0.10$, $df = 221.00$, $t = -9.69$, $p < .001$).
Figure 2

*Mean ratings of perceived robot agency (top) and experience (bottom)*

*Note.* Ratings were collected during the pre-test norming study and were based on the four robots that were later shown in the experimental tasks. Humanoid robots are Pepper and NAO, Lego robots are Gresh and Princess.
Our work in Chapter 2 shows that speakers show biases of mention order when describing dyadic interactions, tending to mention a man first and a woman second reminiscent of effects of androcentrism, or to show a Like Me gender bias to mention a person of their own gender first and a person of a different gender second. To help avoid potential confounding effects of gendered names on mention order, we used names for the robots that were deemed to be gender neutral. Participants provided ratings for how female or male, and feminine or masculine they perceived 6 fictional, robot-related names to be, grading these from very female to very male and very feminine to very masculine using a 9-point Likert scale. We chose the four names with average ratings closest to 5 and were thus rated as the most gender neutral (Cybe, Botz, Mech and Roid), to use for the picture description task. It is unclear to what extent the gendered associations of the names themselves modulated the biases found in Chapter 2, but the use of gender-neutral names for the robots allows us to avoid these potential effects.

4.2.1.3. Materials for experimental tasks

We then conducted the picture description task. Within this, participants completed two experimental tasks, a memory and a language task, followed by a post-study questionnaire. Participants viewed photographs of the same eight figures; four humans (one Black woman and man and one white woman and man) and four robots (two humanoid robots, and two Lego robots), all of which were presented against a white background as in Experiments 1 – 4. The human figures also wore neutral coloured clothing of a similar visual salience. Each figure was again assigned a monosyllabic name: for the humans, these were chosen from the previous experiments (Kate, Beth, Fred and Dave) and were counterbalanced within gender across
participants. Gender neutral robot names were counterbalanced outside of anthropomorphism across participants.

Figure 3

*An example of the memory task matrix used in Study 3*

*Note.* This matrix displayed all eight figures who then appeared individually in the memory task and in pairs in the language task. Lego robots ‘Princess’ and ‘Gresh’ (Lego Bionicle) are on the top row from left to right, humanoid robots Pepper and NAO (Softbank Robotics) are on the bottom row from left to right.
Figure 4

*Example stimuli used in Study 3*

![Figure 4 - Example stimuli](image)

*Note.* These are examples of two interactions across the anthropomorphic condition: humanoid trial “arguing” (left) and Lego trial “touching” (right).

In the memory task, participants viewed portrait photographs of the figures (Figure 3). In the language task, participants saw photographs of the figures executing interactions. Figures were depicted executing ten interactions that could all be described using a symmetrical predicate (Gleitman et al., 1996), such as gesturing irritably (verb: *arguing*) and touching palms (*touching*); see Figure 4. These were the same actions depicted in Experiments 2 and 3. Again, we categorised these symmetrical events into three valence conditions: intimate events (dancing, flirting, kissing, marrying), negative events (arguing, fighting, shouting) and casual events (playing, talking and touching).

Half the photographed interactions showed a human with a humanoid robot (humanoid trials, n=40), and half showed a human with a Lego robot (Lego trials, n=40), see Figure 4. Each human figure appeared on 10 humanoid trials and 10 Lego trials.
Participants were prompted to describe the photographs with a fragment of a symmetrical predicate in a transitive frame (e.g., ...is meeting...), thereby requiring them to mention one figure as the sentence subject and thematic agent, and the other as the sentence object and thematic patient.

Participants’ attitudes towards robots, as they related to human-robot interactions, were measured using the NARS (Nomura et al., 2006) in a post-study questionnaire, with answers given on a 5-point Likert scale from Strongly disagree to Strongly agree.

4.2.1.4. Procedure

We used the same procedure as in Experiments 1, 2 and 4 for both the memory task and language task. As in Experiment 1 and 2, participants viewed an active transitive sentence fragment for 2000 ms and were then prompted to complete the sentence based on the photograph they had just seen, using the names of the figures they had just learnt and typing their description into a text box (e.g., dave is arguing with cybe – Figure 4). We randomised the order of trials.

After the language task, participants completed the NARS and a demographic questionnaire, on which they provided information about their age, first language, current location, gender and race, in addition to answering the question “have you been living in the UK for the last five years?”, to which participants answered yes or no.

This experiment was programmed using JavaScript and the jsPsych library (de Leeuw, J. R., 2015), and data was collected on the Prolific.com platform.
4.2.1.5. Analysis

The study involved both a within-subjects design, in which all participants viewed the same photographs, sentence fragments and anthropomorphism condition, and a between-subjects factor of participant identity, allowing us to investigate differences between participant groups (e.g., women compared to men) and any effects resulting from whether the participant matched the human in the photograph for gender or race.

Consequently, we analysed trials in two ways, firstly with a focus on the animacy bias, predicting that participants would tend to mention the human first and the robot second and that this would be more likely on trials with a larger animacy disparity (i.e., on human/Lego trials than on human/humanoid trials).

We used a similar approach for analyses as in Chapter 2, using mixed effects logistic regressions with mention order as the dependent variable (i.e., if the human was mentioned first) and took an iterative approach to model building for the animacy analysis, which started by determining if a human-first bias was present. We then examined how this interacted with our primary variables of anthropomorphism, participant gender and event valence. Anthropomorphism and participant gender were contrast coded and valence was dummy coded for all analyses. We then entered participants’ scores from each subscale of the NARS to the model as secondary variables, after excluding any individual items that had a low item-total correlation (determined by falling below 0.5 Cronbach's α). Scores on subscale 3 (Negative attitude toward emotional interactions with robots) were reverse-coded.

We also tested whether the human-first bias would be affected by participant age, entering this as a predictor in the main model. Additionally, we tested interactions between participant age and attitudes towards robots. For the Like Me analyses, we ran
a separate model that examined interactions between our participants’ identities (gender and race) and depicted humans’ identities (gender and race), for example, to determine if female participants were more likely to mention the human first when the human was also female, if Black participants were more likely to mention the human first if the human was Black, and so on.

All analyses were performed using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). Models were built, simplified and reported using the same principles as in Experiments 1 – 5. We limited analyses to two-way interactions for the sake of model convergence.

4.2.1.6. Control analyses

Control analyses tested whether the size of any human-first bias changed from the experiment’s start to its end, entering trial number as a predictor in the main model. Lastly, we tested the effect of memory on mention order, examining whether participants were more accurate at remembering the human names over the robot names, as this might be a possible explanation of human-first bias. To do so, we examined data from the first round of the memory task and used a mixed effects model to determine if accuracy was predicted by the relative animacy of the depicted figures (i.e., human versus robot).

4.2.1.7. Exclusions

Rigorous exclusion criteria, which were all preregistered, ensured that participants’ Prolific-reported demographic background matched their self-reported demographic background, and that they completed the experiment correctly. We excluded 4 participants whose reported race did not match with their Prolific records, 1
participant who did not report their gender and 22 participants who reported that they did not speak English as a first language. 18 participants were excluded for not having lived in the UK for the last 5 years. We conducted a familiarity check at the end of the experiment, asking if participants had ever interacted with any of the four robots before in real life and excluded participants who said yes to one or more of the robots (28 total).

Next, we excluded individual trials where participants made response errors (e.g., using incorrect names, verbs, or syntax), comprising 16% of total trials. As in Experiments 1, 2 and 4, trials where participants made a clear typing error that did not signify a figure identification or syntax use error were hand corrected and included. Next, we excluded participants who made response errors on at least 50% of trials (6 participants total).

We further excluded participants who showed a strong left or right orientation bias, naming the left figure or the right figure first on more than 90% – 11 participant total. Mirror-reversing images avoided notable orientation biases across participants (mean orientation bias 0.49% [SD = 0.15]). 2% of trials were excluded for reactions times greater than 3 standard deviations from the participant mean. Some participants were excluded for multiple criteria. Thus, the main analysis included 168 participants (35 Black women, 29 Black men, 56 white women, 48 white men, mean age = 33.7 [SD = 12.68], range 18 – 70).

For the questionnaire analysis, we further excluded participants who routinely selected the same response on over 90% of items, and who did not complete the relevant questionnaire in full. Thus, the NARS analysis was conducted with 152 participants (34 Black women, 24 Black men, 50 white women, 44 white men, mean age = 33.79 [SD = 12.54]). We excluded two items from the NARS analysis for
demonstrating item-total correlations of under 0.5 Cronbach's $\alpha$: one item from subscale 1, “The word “robot” means nothing to me.” (0.28) and one item from subscale 2, “I feel that in the future society will be dominated by robots.” (0.29).

4.2.2. Results

4.2.2.1. Primary analysis

We analysed if participants showed a bias to mention the human figure first over the robot figure, as predicted by the animacy bias, and the extent to which mention order was influenced by anthropomorphism, valence and participant gender. Figure 5 displays these effects.

To do so, we conducted a mixed effects logistic regression of the form Human First ~ (Anthropomorphism + Participant Gender + Valence) $^2 + (1 | \text{Participant}) + (1 | \text{Event})$. The results from this regression are displayed in Table 1.
Figure 5

The mean proportion of trials on which the human was mentioned first

*Note.* Means are for female and male participants, across anthropomorphism and valence conditions. The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot. Error bars represent standard error.
Table 1

Results from the human-first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.17</td>
<td>0.22</td>
<td>9.91</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Anthropomorphism</td>
<td>-0.15</td>
<td>0.05</td>
<td>-2.83</td>
<td>0.00 **</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.09</td>
<td>0.09</td>
<td>-1.01</td>
<td>0.31</td>
</tr>
<tr>
<td>Participant gender</td>
<td>-0.01</td>
<td>0.21</td>
<td>-0.03</td>
<td>0.98</td>
</tr>
<tr>
<td>Anthropomorphism: Valence</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>Anthropomorphism: Participant gender</td>
<td>0.09</td>
<td>0.03</td>
<td>2.73</td>
<td>0.01 **</td>
</tr>
<tr>
<td>Valence : Participant gender</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.72</td>
<td>0.47</td>
</tr>
</tbody>
</table>

**p < .01. ***p < .001.

With these stimuli, participants were overall more likely to mention the human first rather than the robot (M = 0.74 [SD = 0.26]), confirmed by a statistically significant intercept for the regression. Importantly, anthropomorphism had a significant effect on mention order. Contrary to our predictions, participants were more likely to mention the human first when the robot they were interacting with was humanoid rather than Lego (M = 0.74 [SD = 0.27] vs. M = 0.73 [SD = 0.26]).

We also found a significant interaction between anthropomorphism and participant gender; men were more likely to show a human-first bias on trials with human/humanoid dyads rather than on trials with human/Lego dyads (M = 0.77 [SD = 0.24] vs. M = 0.73 [SD = 0.26]). For women, the difference between human/humanoid and human/Lego trials was in the other direction, but also much smaller: women were
slightly more likely to show a human-first bias on human/Lego trials, rather than on human/humanoid trials (M = 0.73 [SD = 0.27] vs. M = 0.72 [SD = 0.29]). Participant gender was not a significant predictor on its own, nor did we find main effects of event valence, or any other interactions.

Next, we analysed the effect of attitudes towards robots, using participants’ scores on the NARS (three subscales: Negative attitude toward interaction with robots; Negative attitude toward the social influence of robots; Negative attitude toward emotional interactions with robots). We added each of the three subscales as scaled variables into an additional model, of the form Human First ~ General Interaction + (Social Influence + Emotional Interactions + Participant Gender) ^2 + Anthropomorphism + (1 | Participant) + (1 | Event). The general interaction subscale did not interact significantly with any of the other fixed effects, so these interactions were removed from the model for convergence, as was the non-significant fixed effect of valence. Results are displayed in Table 2.
Table 2

Results from the human-first linear regression with NARS subscale scores as predictors

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.06</td>
<td>0.26</td>
<td>7.97</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>General interaction</td>
<td>-0.36</td>
<td>0.28</td>
<td>-1.28</td>
<td>0.20</td>
</tr>
<tr>
<td>Social influence</td>
<td>-0.32</td>
<td>0.32</td>
<td>-1.00</td>
<td>0.32</td>
</tr>
<tr>
<td>Emotional interactions</td>
<td>0.37</td>
<td>0.30</td>
<td>1.24</td>
<td>0.22</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.06</td>
<td>0.22</td>
<td>0.26</td>
<td>0.80</td>
</tr>
<tr>
<td>Anthropomorphism</td>
<td>-0.06</td>
<td>0.03</td>
<td>-2.17</td>
<td>0.03 *</td>
</tr>
<tr>
<td>Social influence : Emotional interactions</td>
<td>-0.12</td>
<td>0.22</td>
<td>-0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>Social influence : Participant gender</td>
<td>-0.40</td>
<td>0.31</td>
<td>-1.31</td>
<td>0.19</td>
</tr>
<tr>
<td>Emotional interactions : Participant gender</td>
<td>0.85</td>
<td>0.29</td>
<td>2.98</td>
<td>0.00 **</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001.

We found a significant interaction between participant gender and the emotional interactions measure (subscale 3, see Figure 6); men who had more favourable attitudes towards emotional interactions with robots were less likely to mention the human first. Women showed the opposite effect; those who had more favourable attitudes towards emotional interactions with robots were more likely to mention the human first and the robot second. Neither subscale was a significant predictor of mention order on its own.
Figure 6

Effects of attitudes towards robots on the human-first bias

Note. Lines of best fit represent the mean proportion of trials on which male and female participants mentioned the human first, by overall score on subscale 3 of the NARS (Negative attitude toward emotional interactions with robots). The higher the NARS subscale 3 score, the more positive the attitude.

**Participant age.** We ran a further model, entering participant age as a predictor into the main model. We found a significant main effect of participant age ($\beta = 0.51$, $SE = 0.20$, $z = 2.50$, $p < .05$). Older participants were significantly more likely to mention the human first than younger participants. We also tested age x NARS interactions to determine if younger participants showed differing attitudes towards robots than older participants, however, we did not find that participant age significantly interacted with attitudes towards robots for either of the three surveys.
A control analysis tested whether mention order behaviour changed over the course of the experiment. Trial order was not a significant predictor of the human-first bias. 11

4.2.2.2. Like Me Analyses

Next, we analysed whether participants would be more likely to mention the human first if that figure was also the same gender or the same race as themselves, testing effects of participant gender and race as well as depicted human gender and race and investigating interactions between these variables. These effects are displayed in Figure 7.

To test these effects, we ran a model of the form Human First ~ 1 + (Participant Gender + Human Gender + Human Race) ^2 + Participant Race + (1 | Participant) + (1 | Event). Non-significant interactions with participant race were removed for model convergence, results are displayed in Table 3.

11 Trial Progression. Trial order was entered as a scaled predictor to the main model. Results indicated that trial order was not a significant predictor of the human-first bias (θ = 0.04, SE = 0.03, z = 1.64, p = .10).
Figure 7

Effects of a Like Me bias on the human-first bias

Note. The mean proportion of trials on which participants mentioned the human first, by participant demographic and human demographic. The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot. Error bars represent standard error.
Table 3

Results from the human-first linear regression with participant and human demographics as predictors

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>SE</th>
<th>z</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.10</td>
<td>0.22</td>
<td>9.43</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.00</td>
<td>0.20</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Human gender</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.47</td>
<td>0.64</td>
</tr>
<tr>
<td>Human race</td>
<td>0.01</td>
<td>0.03</td>
<td>0.21</td>
<td>0.84</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.14</td>
<td>0.21</td>
<td>0.66</td>
<td>0.51</td>
</tr>
<tr>
<td>Participant gender : Human gender</td>
<td>0.03</td>
<td>0.03</td>
<td>1.21</td>
<td>0.23</td>
</tr>
<tr>
<td>Participant gender : Human race</td>
<td>0.06</td>
<td>0.03</td>
<td>2.48</td>
<td>0.01 *</td>
</tr>
<tr>
<td>Human gender : Human race</td>
<td>0.07</td>
<td>0.03</td>
<td>2.52</td>
<td>0.01 *</td>
</tr>
</tbody>
</table>

* \( p < .05 \). ** \( p < .001 \).

The interaction between participant gender and depicted human gender was not significant, nor was the interaction between participant race and depicted human race (this interaction was also removed from the model for convergence). Thus, we did not find evidence to support a Like Me gender or race effect on these trials. Neither participant gender, participant race, human gender or human race were significant predictors on their own.

We found a significant interaction between depicted human gender and depicted human race; a human-first bias occurred most often on trials depicting either a Black man (M = 0.75 [SD = 0.26]) or a white woman (M = 0.74 [SD = 0.27]), and least often
on trials with a white man (M = 0.73 [SD = 0.28]) and Black woman (M = 0.73 [SD = 0.27]). We also found a significant interaction between participant gender and human race; male participants were more likely to mention the human first if this figure was Black (M = 0.75 [SD = 0.24]) rather than white (M = 0.74 [SD = 0.26]), whereas female participants were more likely to mention the human first if this figure was white (M = 0.74 [SD = 0.28]) rather than Black (M = 0.72 [SD = 0.28]).

### 4.2.2.3. Memory effects

Lastly, we examined how accurate participants were at remembering the names of the eight photographed figures, analysing their mean trial success during the first round of the memory task. Participants generally demonstrated relatively high accuracy on this task (Black women: M = 0.94 [SD = 0.09]; Black men: M = 0.94 [SD = 0.12]; white women: M = 0.96 [SD = 0.07]; white men: M = 0.94 [SD = 0.10]). To determine if participants were better at memorising the names of the human figures over the names of the robots, we used a mixed effects logistics regression of the form Accuracy ~ (Figure Animacy + Participant Gender + Participant Race)^2 + (1 | Participant) + (1 | Stimulus). Table 4 shows the results from this model.
Table 4

Results from the memory accuracy linear regression

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.40</td>
<td>0.44</td>
<td>7.75</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Figure animacy</td>
<td>-0.78</td>
<td>0.29</td>
<td>-2.68</td>
<td>0.01 **</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.11</td>
<td>0.51</td>
<td>0.22</td>
<td>0.82</td>
</tr>
<tr>
<td>Participant race</td>
<td>-0.07</td>
<td>0.47</td>
<td>-0.14</td>
<td>0.89</td>
</tr>
<tr>
<td>Figure animacy : Participant gender</td>
<td>-0.22</td>
<td>0.30</td>
<td>-0.76</td>
<td>0.45</td>
</tr>
<tr>
<td>Figure animacy : Participant race</td>
<td>0.16</td>
<td>0.30</td>
<td>0.54</td>
<td>0.59</td>
</tr>
<tr>
<td>Participant gender : Participant race</td>
<td>0.32</td>
<td>0.60</td>
<td>0.53</td>
<td>0.59</td>
</tr>
</tbody>
</table>

**\( p < .01 \). ***\( p < .001 \).

Note. Results from a linear regression predicting accuracy in the first round of the memory task from participant and stimulus factors.

We found a significant main effect of figure animacy on memory task accuracy; participants were much better at remembering the names of the human figures, showing greater accuracy on these trials, than the names of the robot figures (\( M = 0.98 \) [SD = 0.07] vs. \( M = 0.92 \) [SD = 0.13]). No interactions were significant between these variables.

4.2.3. Discussion

Results from this experiment show that the animacy bias holds for descriptions of human/robot interactions. Participants in this study reliably showed a human-first
bias to mention the human first and the robot second. We also found an effect of anthropomorphism, but this was unexpectedly in the opposite direction than the one we had initially predicted.

Namely, we had initially predicted that the human-first bias would be strongest on trials with a larger animacy disparity, expecting participants to mention the human first more often on human/Lego trials than on human/humanoid trials. This prediction was made from the assumptions of conceptual accessibility; that entities closer to humanness (such as humanoid robots) will be more accessible than entities further away from humanness (such as the Lego robots) and are therefore more likely to come earlier on in production than their less-animate counterparts (Branigan et al., 2008; Prat-Sala & Branigan, 2000). The results from this experiment indicated that the human-first bias was in fact strongest on human/humanoid trials.

Furthermore, an unexpected significant interaction between anthropomorphism and participant gender suggested that male participants were more likely to mention the human first on human/humanoid trials than female participants, who instead showed little difference between these types of trial. There were two other significant interactions which had no a priori predictions; this time between depicted human gender and depicted human race (participants were more likely to mention the human first if this figure was a Black man or a white woman, rather than a Black woman or a white man), and between participant gender and depicted human race (male participants showed a stronger human-first bias on trials depicting Black people rather than white people, whereas female participants showed a stronger human-first bias on trials depicting white people rather than Black people).

Investigations of how attitudes towards robots may impact a human-first bias revealed one significant interaction, between participant gender and the NARS measure
of attitudes towards emotional interactions with robots. Men with more positive attitudes towards emotional interactions were less likely to mention the human first, whereas women with more positive attitudes towards such interactions with robots were more likely to mention the human first. None of the NARS subscales were significant predictors on their own. Older participants were more likely to mention the human first compared to younger participants. This was unaffected by the attitudes participants of different ages held towards robots.

Lastly, we did not find good evidence for a Like Me bias in this experiment, perhaps due to the overwhelming tendency to put the human first.

4.3. **Experiment 7**

Experiment 7 aimed to replicate the effects from Experiment 6, including the main results but especially the more unexpected findings, to determine if these would hold with a new population and thus warrant further exploration. It also aimed to continue our work testing the Like Me bias, investigating how mention order is influenced by relative social similarity, this time as determined by animacy and inherent accessibility. To do so, we used a narrative manipulation to alter perceptions of different properties of social cognition and ability that are attributable to animacy: agency, experience, empathy and dehumanisation. In doing so, the humans were presented as either more animate or less animate than the robots, and we measured how relative social similarity between these groups and the participants, according to properties of humanness, affected mention order.

Social cognition is dependent on a person’s ability or willingness to consider the mind of another person. We propose that doing so is therefore linked with the theory of
conceptual accessibility, as an entity that has high mind abilities and can therefore execute more actions, express more behaviours and have more experiences will theoretically be more conceptually accessible than an entity that has low mind abilities.

Participants were assigned to two groups to create the narrative condition: human-oriented or robot-oriented. We used narrative in the form of a short story to manipulate inherent accessibility. Participants in the human-oriented condition read a story in which the humans were presented as being more animate than the robots, as is the usual default. In the robot-oriented condition, the robots were presented as more animate than the humans and therefore as more similar to the participants. After reading the story, participants completed the same memory and language tasks as before. They finished by providing ratings of dimensions of mind perception (agency and experience) and empathy in regard to both the human and robot groups, and ratings denoting dehumanisation of the human group. We tested the extent to which presenting the robot group as more human-like than the human group would reverse the human-first bias. We also tested if it was possible to make predictions of mention order based on participants’ scores on the questionnaire, using ratings of relative animacy and dehumanisation as predictors of first mention.

Finally, this new experiment also allowed us to test a further counter-explanation of the original animacy bias – that participants showed a first mention bias for the humans because their human names were easier to memorise and retrieve than the fictional robot names on account of their greater familiarity. Certainly, our analysis of the memory task indicated that participants showed a greater accuracy for recalling the human names. Thus, in this experiment, all figures were given fictional names, that were balanced in familiarity and memorability. This is important for recall purposes, but it may also be important for our animacy manipulation. Human names trigger empathy
for humans and non-human entities (Vaes et al., 2016), so in order to avoid the effect of human names on linguistic accessibility and animacy perceptions, all figures were given fictional names.

4.3.1. Method

We preregistered the experiment on osf.io (https://osf.io/8zxyk), describing the participant sample, exclusion criteria, response coding and intended analyses.

4.3.1.1. Participants

Using Prolific.com, we paid 336 English-speaking adults localised in the UK £4 to take part for 30 minutes. 84 participants reported that they were Black women, 84 reported they were Black men, 84 reported they were white women, and 84 reported they were white men. Mean reported age was 30.78 [SD = 10.71] with a range of 18 – 75. 240 participants (60 from each of the four demographic groups) took part in the human-oriented condition and 96 participants took part in the robot-oriented condition (24 from each demographic group).

To determine sample size, we conducted an effect size analysis with the data from Experiment 6, using the mean proportion of trials on which the human figure was mentioned first. With an effect size of 0.92 and an alpha of less than .05, the power of this study would be 1 with 240 participants in the human-oriented condition. If our manipulation condition (human-oriented/robot-oriented) were to have an effect size of even half this, power would be 99.9% with 240 participants. Furthermore, we required the same number of participants in the human-oriented condition as in Experiment 6 for replication purposes and to effectively investigate the possible effects of name familiarity.
For the robot-oriented condition, a power analysis was conducted in which the 'human-first' effect size of Experiment 6 was halved (from $d = 0.92$ to $d = 0.46$). When using 96 participants, power would be 89% and therefore high enough to detect the animacy bias effect, which was the main analysis for this condition. With these sample sizes, power would be relatively low for our other comparisons of interest (effects of robot agency and experience, participant demographic). In order to adequately power such comparisons, we would require an impossibly large sample size. Therefore, care should be taken in evaluating these comparisons.

### 4.3.1.2. Materials

Participants were assigned to one of two conditions. Those in the human-oriented condition read a short story in which human-like “biologists” from a fictional planet (the same four human figures as in Experiment 6) are developing robot-like “clones” (the same four robot figures as in Experiment 6), in order to spy on a bio-metallic alien population. Participants in the robot-oriented condition read the same short story, with some crucial differences. Here, the same four robot figures as before appeared as bio-metallic, robot-like “biologists” on a fictional planet, who are developing human-like “clones” represented by the same four human figures as before, in order to spy on a human-like alien population (Figure 8). This condition was designed to anthropomorphise the robot group and dehumanise the human group.
Note. This page introduced the robot group as protagonists. In the human-oriented condition, the human group was introduced as protagonists. The image showed a NASA control room (without the logo) and described Teran as populated by “fleshy, sentient beings”.

Both short stories manipulated perceptions of the more animate group’s mind using cues from Gray et al.’s distinctions of experience and agency interspersed within the narrative. The narrative sought to specifically manipulate perceptions of this group’s experience as related to personality, joy, pride, consciousness, pleasure, fear and desire. Each biologist had a short description detailing their personality presented along with their portrait, and each description then included two details from Gray et al. (2007) that indicated other features of experience. For example:
Biologist Roid is an award-winning researcher whose speciality lies in developmental physiology. Roid also loves [experience feature: joy] to compose music and knows [experience feature: consciousness] how important it is for scientists to have multiple creative outlets.

The final part of the short story included information to suggest the biologists are capable of fear and desire, for example:

The Biologists are nervous [experience feature: fear] about the assessment... Most of all, the Biologists want [experience feature: desire] to help protect their beloved planet from attacks.

This part of the story also included several indicators of the biologists’ agency (also using Gray et al., 2007), namely communication, planning, thought and morality. For example:

After working together for months [agency: communication], the Biologists have reached the first testing stage of this project: they have grown some prototype Spy-Bods, which they have trained to mimic actions with a partner... they think [agency: thought] their Spy-Bods are ready. They are planning [agency: planning] to impress the Space Defence team by interacting with the Spy-Bods themselves.

By describing the less animate group as spy clones, or ‘Spy-Bods’, we intended to dehumanise the human figures in the robot-oriented condition.

After participants had read the short story, they then moved onto the memory task and then the language task as in Experiment 6. This time, both names the participants used were fictional names. This was achieved using the four robot names gathered during the norming study and by modifying the four human names (e.g., Dave became Daxe, Beth became Neth). Names were counterbalanced across human and robot groups, so that humans and robots were both assigned either robot-related
fictional names or modified human names. Again, participants used these names in their descriptions (e.g., “daxe is arguing with cybe”).

Upon finishing the critical trials, participants completed a questionnaire, devised to check if our narrative manipulation had successfully changed the perceived animacy of the robot and human groups. We also used the ratings given on this questionnaire as predictors of mention order. Perceptions of agency and experience were measured using items from Gray et al. (2007) and participants’ empathy towards these figures was measured using items from Seo et al. (2015) and an original item. We collected ratings of dehumanisation of the human figures using items from Harris and Fiske (2011). Participants finished by completing the same robot familiarity check and demographics questionnaire as before.

4.3.1.3. Analysis

We conducted the same analyses as in Experiment 6 for both conditions together and separately, to determine if a human-first bias was present overall and in both conditions, and if it was modulated by the animacy disparity or valence. We also investigated possible Like Me effects on mention order. Additionally, analyses involved comparing the robot-oriented and human-oriented conditions at the group level, to determine if the human-first bias weakened or even reversed when the robot figures were presented as the most animate group.

We used the ratings given in the post-study questionnaire to determine if the human-first bias could be predicted on an individual level by scores of perceived agency, perceived experience, participant empathy and dehumanisation. We also analysed these ratings to determine the effectiveness of the narrative manipulation, comparing relative animacy as measured by agency, experience and empathy scores.
between the human and robot groups, within narrative condition. Previous research has found differences between women and men in their capacity for mentalising; studies have argued that women are better at showing empathy than men (e.g., Abu-Akel & Bo, 2013; Baron-Cohen & Wheelwright, 2004). Thus, we also tested whether participant gender modulated the degree to which our narrative manipulation affected mention order.

Replication analyses were also carried out to help identify significant effects from Experiment 6 that should be considered as reliable behaviours, as were control analyses including trial progression effects and investigations of name memory. The latter was important to rule out linguistic accessibility as the main explanation for the strong human-first bias in Experiment 6.

4.3.1.4. Exclusions

Exclusion criteria were all preregistered. We excluded 8 participants whose reported race did not match with their Prolific records, 3 whose reported gender did not match and 14 participants who reported that English was not their first language. 28 participants were excluded for not having lived in the UK for the last 5 years. We conducted a familiarity check at the end of the experiment, asking if participants had ever interacted with any of the four robots before in real life and excluded participants who said yes for one or more of the robots (31 total). We excluded individual trials where participants made response errors (e.g., using incorrect names, verbs, or syntax), comprising 26% of total trials. Then, we excluded participants who made response errors on at least 50% of trials – 18 participants total. We further excluded participants who showed a strong orientation bias – 8 participants in total. Mirror-reversing images avoided notable orientation biases across participants (mean orientation bias = 0.50 [SD
2% of trials were excluded for reactions times greater than 3 standard deviations from the participant mean.

Some participants were excluded for multiple criteria. Thus, the main analysis included 165 participants in the human-oriented condition (43 Black women, 32 Black men, 51 white women and 39 white men, mean age = 31.33 [SD = 11.19], range 18 – 75) and 72 participants in the robot-oriented condition (21 Black women, 12 Black men, 20 white women and 19 white men, mean age = 31.6 [SD = 10.53], range = 18 – 59).

An important element of our analyses was to test if our narrative manipulation was successful at influencing perceptions of mind, empathy, animacy and dehumanisation. To do so, we firstly identified which items in each sub-scale had a low total-item correlation (Cronbach’s Alpha of under 0.50) and should therefore be removed from analysis. Consequently, we removed the following 5 items: Agency item Memory (“To what extent are Biologists/Spy-Bods capable of remembering things”, Spy-Bod rating for human-oriented: $\alpha = 0.44$, for robot-oriented: $\alpha = 0.37$); Experience item Rage (“To what extent are Biologists/Spy-Bods capable of experiencing violent or uncontrolled anger?”, Biologist rating for human-oriented: $\alpha = 0.43$); Empathy item Worry (“How worried are you for the Biologists/Spy-Bots?”, Spy-Bod robot-oriented: $\alpha = 0.47$); Dehumanisation general traits item Familiar (“How familiar to you are the Biologists/Spy-Bots?”, human-oriented: $\alpha = 0.41$) and Dehumanisation story traits item Animal (“I felt like the Biologists/Spy-Bots in the description lacked self-restraint, like an animal”, robot-oriented: $\alpha = 0.10$).
4.3.2. **Results**

4.3.2.1. **Analysis of manipulation check**

First, we tested if our narrative manipulation was successful at altering conceptual accessibility and influencing perceptions animacy of the human and robot figures. To do so, we firstly compared mean ratings of humans and robots from three sub-scales of the post-study questionnaire across narrative conditions. We expected participants in the human-oriented condition to give the human group higher ratings of animacy than the robot group, and we expected participants in the robot-oriented condition to give the human group lower ratings of animacy than the robot group. These sub-scales tested the following: perceived agency and perceived experience, both using items from Gray et al. (2007), and empathy using items from Seo et al. (2015) and an original item. Figure 9 displays these differences.

We ran a linear mixed model using the `lmer` function from `lmerTest`, investigating interactions between the narrative condition (human-oriented/robot-oriented), rated figure (human group/robot group) and the scores from these ratings. This was of the form `Score ~ Narrative Condition * Sub-Scale * Rated Figure + (1 + Sub-Scale + Rated Figure | Participant)`. Narrative condition and rated figure were both contrast coded. The results of this model are displayed in Table 5.
Figure 9

Mean ratings of animacy for humans and robots

Note. The mean ratings of agency, experience and empathy for human and robot figures across the narrative condition. Error bars represent standard error.
Table 5

Results from the manipulation check linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>60.37</td>
<td>0.94</td>
<td>234.99</td>
<td>64.21</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition</td>
<td>0.85</td>
<td>0.94</td>
<td>234.99</td>
<td>0.90</td>
<td>0.37</td>
</tr>
<tr>
<td>Empathy</td>
<td>-14.40</td>
<td>1.44</td>
<td>235.01</td>
<td>-10.03</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Experience</td>
<td>-5.06</td>
<td>0.87</td>
<td>234.99</td>
<td>-5.82</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Rated figure</td>
<td>-8.35</td>
<td>1.02</td>
<td>295.13</td>
<td>-8.22</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition : Empathy</td>
<td>-0.21</td>
<td>1.44</td>
<td>235.01</td>
<td>-0.14</td>
<td>0.89</td>
</tr>
<tr>
<td>Narrative condition : Experience</td>
<td>-0.23</td>
<td>0.87</td>
<td>234.99</td>
<td>-0.27</td>
<td>0.79</td>
</tr>
<tr>
<td>Narrative condition : Rated Figure</td>
<td>10.74</td>
<td>1.02</td>
<td>295.13</td>
<td>10.58</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Empathy : Rated figure</td>
<td>4.41</td>
<td>0.70</td>
<td>8054.00</td>
<td>6.31</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Experience : Rated figure</td>
<td>-6.18</td>
<td>0.51</td>
<td>8054.00</td>
<td>-12.12</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition : Empathy : Rated Figure</td>
<td>-4.42</td>
<td>0.70</td>
<td>8054.00</td>
<td>-6.32</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition : Experience : Rated Figure</td>
<td>2.91</td>
<td>0.51</td>
<td>8054.00</td>
<td>5.72</td>
<td>0.00 ***</td>
</tr>
</tbody>
</table>

***p < .001.

The interaction between narrative condition and rated figure was significant. This suggests that, depending on the narrative condition they had been assigned to, participants rated humans and robots differently for these properties of animacy. As expected, participants in the human-oriented condition gave the human figures higher ratings of animacy than the robots (Agency human M = 78.61 [SD = 15.29], Agency robot M = 40.44 [SD = 21.90]; Experience human M = 82.88 [SD = 16.58], Experience robot M = 26.50 [SD = 20.71]; Empathy human M = 55.59 [SD = 23.04], Empathy
robot M = 35.06 [SD = 24.91]). Importantly, however, participants in the robot-oriented condition gave more similar ratings of animacy for the humans and robots (Agency human M = 58.82 [SD = 22.86], Agency robot M = 63.61 [SD = 23.46]; Experience human M = 56.80 [SD = 26.24], Experience robot M = 55.05 [SD = 23.87]; Empathy human M = 44.22 [SD = 25.20], Empathy robot M = 49.00 [SD = 26.97]), contrary to predictions.

Next, we analysed this at sub-scale level, to determine if this interaction was significant for all three measures of animacy. We ran three individual models, taking subsets of the data by agency scores, experience scores and empathy scores separately, and analysing the interaction between narrative condition and rated figure. We found this interaction to be significant in all three models (agency model $\beta = 10.74$, $SE = 1.08$, $df = 234.99$, $t = 9.95$, $p < .001$; experience model $\beta = 13.66$, $SE = 1.17$, $df = 235$, $t = 11.70$, $p < .001$; empathy model $\beta = 6.32$, $SE = 1.05$, $df = 235$, $t = 6.00$, $p < .001$). Interactions between condition and experience, and condition and empathy were non-significant, indicating that the effect of narrative condition on scores was roughly the same for all three measures of animacy. The significant effect of rated figure indicated that animacy scores differed depending on whether the participant was rating the human group or the robot group.

Dehumanisation. Next, we analysed if participants showed a dehumanising response to our narrative manipulation and were therefore more likely to dehumanise the human figures in the robot-oriented condition than in the human-oriented condition. We compared mean ratings of dehumanisation towards the human figures between the human-oriented and robot-oriented narrative conditions, expecting that participants in the robot-oriented condition showed greater dehumanisation than participants in the
human-oriented condition. These ratings were collected from two dehumanisation sub-scales in the post-study questionnaire (General Traits and Story Traits) using items from Harris and Fiske (2011). Effects are depicted in Figure 10. To test this, we ran a linear regression of the form Score ~ Narrative Condition * Sub-Scale + (1 + Sub-Scale | Subject). Results are displayed in Table 6.

**Figure 10**

*Mean dehumanisation scores*

Note. The mean ratings on the dehumanisation measure for human figures across the narrative condition. Error bars represent standard error.
**Table 6**

*Results from the dehumanisation linear regression*

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>df</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>72.34</td>
<td>1.24</td>
<td>235</td>
<td>58.12</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition</td>
<td>-13.10</td>
<td>2.26</td>
<td>235</td>
<td>-5.80</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Sub-scale</td>
<td>-4.92</td>
<td>1.18</td>
<td>235</td>
<td>-4.15</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition : Sub-scale</td>
<td>-1.76</td>
<td>2.15</td>
<td>235</td>
<td>-0.82</td>
<td>0.41</td>
</tr>
</tbody>
</table>

***p < .001.

Crucially, a significant effect of narrative condition indicated that scores were much lower in the robot-oriented condition than in the human-oriented condition. As expected, participants in the robot-oriented condition were more likely to dehumanise the human figures than participants in the human-oriented condition. Ratings were lower in the robot-oriented condition than the human-oriented condition for both sub-scales (robot-oriented general traits M = 59.24 [SD = 19.89], story traits M = 52.57 [SD = 17.38] vs. human-oriented general traits (M = 72.34 [SD = 13.96], story traits M = 67.42 [SD = 16.58]). Figure 10 displays these trends. A significant main effect of sub-scale indicated that scores were lower on the story traits measure than the general traits measure. The interaction between narrative condition and sub-scale was not significant, indicating that the difference in scores between the robot-oriented and human-oriented condition was similar for both sub-scales.
4.3.2.2. Analysis of the narrative manipulation on the human-first bias

Next, we analysed if participants showed a bias to mention the human figure first over the robot figure overall, and crucially if they were less likely to do so in the robot-oriented condition than in the human-oriented condition. Through our other fixed effects, we investigated if mention order was affected by anthropomorphism (i.e., whether trials depicted human/humanoid pairs or human/Lego pairs), by the valence of the events (i.e., intimate trials, casual trials and negative trials) and by participant gender (women/men). Figures 11 and 12 display these effects.

Figure 11

The mean proportion of trials on which the human was mentioned first

Note. Means are for female and male participants in both narrative conditions and across the anthropomorphism condition. The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot. Error bars represent standard error.
Figure 12

*The mean proportion of trials on which the human was mentioned first, by valence*

![Bar chart showing the mean proportion of trials on which the human was mentioned first, by valence.](image)

*Note.* Means across both narrative conditions and event valence. The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot. Error bars represent standard error.

We conducted a mixed effects logistic regression of the form Human First ~ Narrative Condition * Valence + Anthropomorphism * Participant Gender + (1 + Valence | Subject) + (1 | Event). Non-significant interactions were removed for model convergence. The results from this regression are displayed in Table 7.
Table 7

Results from the human first linear regression

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.63</td>
<td>0.18</td>
<td>8.89</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition</td>
<td>-0.86</td>
<td>0.18</td>
<td>-4.84</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Anthropomorphism</td>
<td>-0.09</td>
<td>0.02</td>
<td>-3.99</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.09</td>
<td>0.05</td>
<td>-1.72</td>
<td>0.09 *</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.22</td>
<td>0.17</td>
<td>1.31</td>
<td>0.19</td>
</tr>
<tr>
<td>Narrative condition : Valence</td>
<td>-0.04</td>
<td>0.03</td>
<td>-1.33</td>
<td>0.18</td>
</tr>
<tr>
<td>Anthropomorphism: Participant gender</td>
<td>0.05</td>
<td>0.02</td>
<td>2.26</td>
<td>0.02 *</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001. * = marginally significant.

A statistically significant intercept for the regression indicated that participants were overall more likely to mention the human first rather than the robot (M = 0.71 [SD = 0.28]). Crucially, we found a significant main effect of narrative condition; the human-first bias was significantly less likely in the robot-oriented condition than in the human-oriented condition (M = 0.57 [SD = 0.28] vs. M = 0.78 [SD = 0.26]). A main effect of anthropomorphism indicated that participants were more likely to mention the human first when the human was paired with a humanoid robot than with a Lego robot (M = 0.72 [SD = 0.28] vs. M = 0.70 [SD = 0.29]). Event valence was a marginally significant predictor of mention order.

Lastly, we replicated the participant gender and anthropomorphism interaction from Experiment 6. For male participants, the human-first bias was more likely to occur
on trials with human/humanoid dyads than on trials with human/Lego dyads (M = 0.70 [SD = 0.28] vs. M = 0.66 [SD = 0.30]), whereas for female participants the difference was negligible between humanoid trials and Lego robot trials (M = 0.74 [SD = 0.28] vs. M = 0.73 [SD = 0.29]). Participant gender was not a significant predictor on its own.

To determine if the human-first bias was significant in both narrative conditions, we ran additional models using data from only the human-oriented condition and only the robot-oriented condition separately. We found significant intercepts for both models (human-oriented $p < .001$; robot-oriented $p < .01$); participants in both narrative conditions were significantly more likely to mention the human first than the robot.

Thus, while our narrative manipulation significantly reduced the likelihood of a human-first bias when robots were presented with greater perceived mind than the human figures, it did not outright reverse this bias.

### 4.3.2.3. Analysis of manipulation check ratings on the human-first bias

Next, we used ratings from the post-study questionnaire to create new measures of relative animacy – namely relative agency, experience and empathy between the human and robot figures. We calculated this by subtracting scores for the robot group from scores of the human group. We then used relative agency, experience and empathy as scaled predictors of the human-first bias in a new model and investigated their interactions with another participant variable: participant gender. This model was of the form Human First ~ (Participant gender + Relative Agency + Relative Experience + Relative Empathy)$^2$ + (1 | Subject). Random effects by item were removed for model convergence. Results are displayed in Table 8.
Table 8

Results from the human-first regression with relative animacy scores as predictors

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.03</td>
<td>0.21</td>
<td>9.76</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.22</td>
<td>0.17</td>
<td>1.30</td>
<td>0.19</td>
</tr>
<tr>
<td>Relative agency</td>
<td>-0.15</td>
<td>0.27</td>
<td>-0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Relative experience</td>
<td>0.42</td>
<td>0.30</td>
<td>1.41</td>
<td>0.16</td>
</tr>
<tr>
<td>Relative empathy</td>
<td>0.41</td>
<td>0.21</td>
<td>2.01</td>
<td>0.04 *</td>
</tr>
<tr>
<td>Participant gender : Relative agency</td>
<td>0.59</td>
<td>0.27</td>
<td>2.18</td>
<td>0.03 *</td>
</tr>
<tr>
<td>Participant gender : Relative experience</td>
<td>-0.57</td>
<td>0.28</td>
<td>-2.03</td>
<td>0.04 *</td>
</tr>
<tr>
<td>Participant gender : Relative empathy</td>
<td>-0.02</td>
<td>0.21</td>
<td>-0.07</td>
<td>0.94</td>
</tr>
<tr>
<td>Relative agency : Relative experience</td>
<td>-0.14</td>
<td>0.18</td>
<td>-0.81</td>
<td>0.42</td>
</tr>
<tr>
<td>Relative agency : Relative empathy</td>
<td>0.01</td>
<td>0.27</td>
<td>0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>Relative experience : Relative empathy</td>
<td>-0.04</td>
<td>0.27</td>
<td>-0.15</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*p < .05. ***p < .001.

From this analysis, we found a significant main effect of relative empathy; participants who showed greater empathy towards the human group compared to the robot group were the most likely to demonstrate a human-first bias. Figure 13 also indicates that, in the case of the robot-oriented condition, participants who showed greater empathy towards the robot group compared to the human group were more likely to mention the robot first.
Figure 13

Effects of relative empathy on the human-first bias

Note. The effect of participants’ relative empathy scores from the post-study questionnaire on the human-first bias, across the narrative condition. A score of above 0 indicates the participant gave the human group a higher empathy score than the robot group. A score of below 0 indicates an empathy score in favour of the robot group. The dotted line represents chance at 50% - above the line represents a mention order bias in favour of the human, below represents a bias in favour of the robot.

Results from this model also revealed two significant interactions: between participant gender and relative agency, and participant gender and relative experience. Figure 14 shows the former interaction; in the human-oriented condition, participants showed a human-first bias on around 75% of trials regardless of agency ratings. In the robot-oriented condition, however, participants who gave the human group higher ratings of agency than the robot group were significantly more likely to mention the
human first than the robot. For male participants, however, if they had given the robot group higher ratings of agency than the human group, they were also more likely to mention the robot first.

**Figure 14**

*Effects of relative agency on the human-first bias*

![Graph showing effects of relative agency on the human-first bias.](image)

**Note.** Participants’ relative agency scores from the post-study questionnaire were used as predictors of the human-first bias. This is presented across the narrative condition and participant gender. A score of above 0 indicates the participant gave the human group a higher agency score than the robot group. A score of below 0 indicates an agency score in favour of the robot group. The dotted line represents chance at 50% - above the line represents a mention order bias in favour of the human, below represents a bias in favour of the robot.
Figure 15 shows the latter interaction; in the human-oriented condition, participants generally showed a strong human-first bias, regardless of experience ratings. However, this bias was stronger for male participants who gave the human group much higher ratings of experience than the robot group. In the robot-oriented condition, male participants who gave the human group higher ratings of experience than the robot group were significantly more likely to mention the human first than the robot. If they had given the robot group higher ratings of agency than the human group, they were more likely to mention the robot first. Female participants showed the opposite direction of behaviour – they generally showed a human-first bias, but this increased in strength as they rated the robot group as having greater experience than the human group.

Next, we investigated scores on the two dehumanisation subscales, general traits and story traits, as predictors of the human-first bias along with narrative condition and participant demographic. To do so, we ran a model of the form Human First ~ (Narrative Condition + Participant Gender + General Traits)^2 + Story Traits + (1 | Participant). Non-significant interactions with scores of story traits and random effects by item were also removed for convergences. Results from this analysis are displayed in Table 9.

As in the main model, the intercept and narrative condition were both significant. We also found a significant predictor in scores of story traits; participants were less likely to mention the human first the more they had dehumanised the human group, indicated by lower scores on this sub-scale.
Figure 15

*Effects of relative experience on the human-first bias*

Note. Participants’ relative experience scores from the post-study questionnaire were used as predictors of the human-first bias. This is presented across the narrative condition and participant gender. A score of above 0 indicates the participant gave the human group a higher experience score than the robot group. A score of below 0 indicates an experience score in favour of the robot group. The dotted line represents chance at 50% - above the line represents a mention order bias in favour of the human, below represents a bias in favour of the robot.
Table 9

*Results from the human-first regression with dehumanisation scores as predictors*

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.73</td>
<td>0.19</td>
<td>9.10</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Narrative condition</td>
<td>-0.74</td>
<td>0.19</td>
<td>-3.82</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.15</td>
<td>0.18</td>
<td>0.85</td>
<td>0.40</td>
</tr>
<tr>
<td>General traits</td>
<td>-0.22</td>
<td>0.21</td>
<td>-1.06</td>
<td>0.29</td>
</tr>
<tr>
<td>Story traits</td>
<td>0.47</td>
<td>0.21</td>
<td>2.22</td>
<td>0.03 *</td>
</tr>
<tr>
<td>Narrative condition: Participant gender</td>
<td>-0.30</td>
<td>0.19</td>
<td>-1.61</td>
<td>0.11</td>
</tr>
<tr>
<td>Narrative condition: General traits</td>
<td>0.27</td>
<td>0.17</td>
<td>1.52</td>
<td>0.13</td>
</tr>
<tr>
<td>Participant gender: General traits</td>
<td>-0.49</td>
<td>0.18</td>
<td>-2.78</td>
<td>0.01 **</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001.

Scores on the general traits sub-scale showed a significant interaction with participant gender; the more male participants had dehumanised the human group on this sub-scale, the less likely they were to mention the human first. Figure 16 also indicated that, in the robot-oriented condition, male participants were more likely to mention the robot first if they had given high dehumanisation scores on the general traits sub-scale. Female participants, on the other hand, were slightly less likely to mention the human first if they had given the human group high humanisation scores on this sub-scale. This was most apparent in the human-oriented condition. In both narrative conditions, dehumanisation scores of female participants did not appear to predict a reversal of the human-first bias.
**Figure 16**

*Effects of dehumanisation (general traits) on the human-first bias*

![Graph showing the effect of dehumanisation on the human-first bias](image)

*Note.* The effect on the human-first bias of participants’ scores on the general traits subscale of the dehumanisation measure from the post-study questionnaire, across the narrative condition and participant gender. A lower score indicates greater dehumanisation. The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot.

The interactions between narrative condition and general traits, and narrative condition and story traits were non-significant. This indicates that when an effect of dehumanisation on mention order appeared, it was not driven by the narrative manipulation. Lastly, the interaction between narrative condition and participant gender was also non-significant.
Participant Age. We entered scaled participant age as a predictor in the main model to investigate the effects of participant age. It was not a significant predictor of mention order ($\beta = 0.25, SE = 0.18, z = 1.35, p = .18$). In Experiment 6, participant age was a significant predictor of the human-first bias, with older participants more likely to mention the human first than younger participants. To see if this replicated with the human-oriented data, we ran an additional model with data from this condition only. The effect of participant age was not significant ($\beta = 0.08, SE = 0.21, z = 0.38, p = .71$).

A control analysis tested whether mention order behaviour changed over the course of the experiment, however trial order was not a significant predictor. However, we did find a significant interaction between trial number and condition ($\beta = 0.05, SE = 0.02, z = 2.34, p < .05$; see Figure 17). Over the course of the experiment, participants in the robot-oriented condition showed an increase in their human-first bias, whereas participants in the human-oriented condition showed a very slight decrease.

---

12 Trial Progression. As in Experiment 6, trial order was entered as a scaled predictor to the main model. Results indicated that trial order was not a significant predictor of the human-first bias ($\beta = 0.01, SE = 0.02, z = 0.40, p = .69$). This was also the case in models using either data from the human-oriented condition only, or data from the robot-oriented condition only.
**Figure 17**

*Effects of trial progression on the human-first bias*

*Note.* Lines of best fit representing the mean proportion of trials on which participants mentioned the human first, by trial progression and narrative condition. The dotted line represents chance at 50% - above the line represents a mention order bias in favour of the human, below represents a bias in favour of the robot.

4.3.2.4. **Like Me Analyses**

As in Experiment 6, we analysed whether participants would be more likely to mention the human first if that figure was also the same gender or the same race as themselves, testing effects of participant gender and race as well as depicted human gender and race, and investigating interactions between the variables. These effects are displayed in Figure 18.
Figure 18

*Like Me effects on the human-first bias*

Note. Means are across participant and human demographic (gender and race). The dotted line represents chance at 50% - above the line represents a bias in favour of the human, below represents a bias in favour of the robot. Error bars represent standard error.

To do so, we ran a model with data from the human-oriented condition only, as a replication effort to compare against Experiment 6. This was of the form Human First ~ (Participant Gender + Human Gender + Human Race)^2 + Participant Race + (1 | Participant). Non-significant interactions with participant race and random effect by item were removed for model convergence. Results are displayed in Table 10.
### Table 10

Results from the human-first linear regression with participant and human demographics as predictors (human-oriented only)

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.41</td>
<td>0.31</td>
<td>7.73</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>Participant gender</td>
<td>0.33</td>
<td>0.21</td>
<td>1.58</td>
<td>0.12</td>
</tr>
<tr>
<td>Human gender</td>
<td>-0.15</td>
<td>0.08</td>
<td>-1.85</td>
<td>0.06 *</td>
</tr>
<tr>
<td>Human race</td>
<td>-0.06</td>
<td>0.08</td>
<td>-0.73</td>
<td>0.46</td>
</tr>
<tr>
<td>Participant race</td>
<td>0.25</td>
<td>0.41</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td>Participant gender : Human gender</td>
<td>-0.02</td>
<td>0.06</td>
<td>-0.42</td>
<td>0.68</td>
</tr>
<tr>
<td>Participant gender : Human race</td>
<td>0.00</td>
<td>0.06</td>
<td>-0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>Human gender : Human race</td>
<td>0.20</td>
<td>0.11</td>
<td>1.73</td>
<td>0.08 *</td>
</tr>
</tbody>
</table>

***p < .001, * = marginally significant.

Out of these fixed effects, only human gender, which was non-significant in Experiment 6, showed significance and this was marginal. The interaction between human gender and human race, which was significant in Experiment 6, was only marginally significant here. As in the previous experiment, we did not find any other significant main effects for participant gender, participant race or human race. Unlike in Experiment 6, the interaction between participant gender and human race was not significant.
4.3.2.5. Memory effects

As in Experiment 6, we examined how successful participants were with remembering the names of the eight photographed figures, analysing their mean trial accuracy during the first round of the memory task. Again, participants demonstrated relatively high accuracy on this task (Black women: $M = 0.94 \ [SD = 0.10]$; Black men: $M = 0.93 \ [SD = 0.12]$; white women: $M = 0.96 \ [SD = 0.07]$; white men: $M = 0.95 \ [SD = 0.09]$). We conducted a mixed effects logistic regression of the form $\text{Accuracy} \sim \text{Condition} \ast \text{Figure Animacy} + \text{Participant Gender} + \text{Participant Race} + (1 | \text{Participant}) + (1 | \text{Stimulus})$, with non-significant interactions removed for model convergence.

Results from this model revealed that, despite learning fictional names for both the human and the robot figures, participants were significantly better at remembering the names of the human figures than the robot figures, as in Experiment 6. This was indicated by a significant main effect of figure animacy ($\beta = -0.22, SE = 0.11, z = -1.97, p < .05$). However, this difference was much smaller than in the previous experiment (Experiment 6: $M = 0.98 \ [SD = 0.07]$ vs. $M = 0.92 \ [SD = 0.13]$ vs. Experiment 7 human names $M = 0.96 \ [SD = 0.10]$, robot names $M = 0.94 \ [SD = 0.12]$), suggesting an effectiveness for standardising the names across human and robot figures.

The fixed effect of narrative condition was not significant ($\beta = -0.06, SE = 0.20, z = -0.31, p = .76$); accuracy in the memory task did not notably differ between the human-oriented and robot-oriented narrative manipulation. Importantly, the interaction between figure animacy and narrative condition was not significant ($\beta = 0.08, SE = 0.10, z = 0.82, p = .41$), suggesting that participants were better at remembering the human names regardless of the narrative manipulation they received. This is noteworthy as it shows that, even in the robot-oriented condition where participants were exposed to the robot figures’ names proportionally more than the human figures’ names,
participants found it easier to recall names if they were associated with the human figures. Participant gender ($\beta = 0.07$, $SE = 0.13$, $z = 0.53$, $p = .59$) and participant race ($\beta = 0.20$, $SE = 0.13$, $z = 1.55$, $p = .12$) were not significant predictors of accuracy on this memory task, either.

4.3.3. Discussion

The main effect of narrative condition indicated that participants in the robot-oriented condition, who had been influenced to view the robots as more animate than the humans, were far less likely to mention the human first than participants in the human-oriented condition, who had been influenced to view the humans as the more animate group. Thus, our narrative manipulation was successful in modulating participants’ animacy perceptions of humans and robots. However, participants were overall more likely to mention the human first than the robot, and this was true for both the human-oriented condition and robot-oriented condition. Thus, we were not able to fully reverse the human-first bias with this manipulation.

When testing the influence of relative animacy ratings, we found a significant main effect of relative empathy. Participants were less likely to mention the human first if they showed greater empathy towards the robots than towards the humans. In the robot-oriented condition, this also manifested as a robot-first bias. We also found that dehumanisation modulated mention order. Participants who had shown greater dehumanisation of the human group were also less likely to demonstrate a human-first bias. This was apparent for both narrative conditions, and the effect was stronger on the story traits measure than the general traits measure.

Mention order in this experiment was again affected by anthropomorphism; participants were more likely to show a human-first bias on human/humanoid trials than
human/Lego trials. We replicated the interaction between anthropomorphism and participant gender. Lastly, we investigated if a Like Me bias would manifest on this task, as we have found in Chapter 2, however we did not find evidence that participants were more likely to mention the human first if they were the same gender or race as the participant.

4.4. General Discussion

This study investigated the multi-faceted nature of animacy, using distinctions of animacy between humans and robots as a property that signifies relative social similarity with a speaker. We began, in Experiment 6, by testing the strength of the animacy bias when speakers are describing human-robot dyadic interactions. Here, we aimed to show that mention order could be predicted by greater similarity between the speaker and a group of humans, compared to the speaker and a group of robots, in the form of a robust human-first bias. We varied animacy further using anthropomorphism, by showing interactions between humans and humanoid robots and humans and Lego robots. In doing so, we predicted that the human-first bias would be stronger on trials with a larger animacy disparity between the two figures (i.e., on human/Lego trials rather than on human/humanoid trials). We also investigated effects of attitudes towards robots, using items from the NARS (Nomura et al., 2006) to test if participants’ mention order would be modulated by their attitudes towards interactions with robots.

In Experiment 7, we aimed to replicate the main findings from the previous experiment and demonstrate how the properties that make an entity more animate or less animate than another are flexible and can be altered. To do so, we varied the inherent accessibility of the humans and robots, using a narrative manipulation to indicate that one group was more animate than the other. We tested if we could
modulate the human-first bias by anthropomorphising the robots and dehumanising the humans, perhaps even to the extent that participants would mention the robot first instead. In acknowledgment of the complexity of animacy, we investigated the extent to which perceptions of four different properties of animacy could predict mention order, including agency, experience, empathy and dehumanisation. We predicted that the likelihood of participants mentioning the robot first could be predicted by the extent to which participants perceived the robots as more animate than the humans. In both Experiment 6 and 7, we also explored effects of valence, participant gender, participant age, the Like Me bias and memory for names.

4.4.1. Anthropomorphism and the animacy bias

In both experiments, we found a robust human-first bias. However, in Experiment 7, the human-first bias was much weaker when participants had been induced to view the robot group as more animate than the human group. Our results from this experiment show that mention order can be driven by a proximity to human-like mind, presenting interesting implications for the theory of conceptual accessibility. These findings suggest that inherent conceptual accessibility is flexible, not just derived conceptual accessibility, and that it can be altered through narrative.

Interestingly, we found an effect of anthropomorphism in both experiments. However, this effect was in the opposite direction to what we had initially predicted. The two humanoid robots presented were closer to humanness in physical characteristics than the Lego figures and were also rated as having higher perceived mind than the Lego robots in the norming study. Therefore, they would theoretically appear higher on hierarchies of conceptual accessibility and animacy than the Lego robots, which would suggest that a speaker should be more likely to mention them first
than the Lego robots (Bock & Warren, 1985; Bock, 1986; McDonald et al., 1993; Aissen, 2003; Harris, 1978). In fact, in both experiments we found that participants mentioned the Lego robots first more often than they mentioned the humanoid robots first. Consequently, these results suggest that the current explanation of the animacy hierarchy for mention order biases is perhaps too simplistic.

It is possible that the speakers in this study were influenced by bottom-up effects of perceptual salience, which might explain the unexpected anthropomorphism finding. The Lego robots were both colourful and aesthetically unusual in their toy-like appearances compared to the mostly white and smooth humanoid robots. We might then consider that speakers mentioned the Lego robots first more often than the humanoid robots because the former were more visually salient than the latter.

A higher-level consideration for this replicated anthropomorphism effect is that the likelihood of an animacy or human-first bias may be predicated on the relative animacy between the two entities. We might conclude from this study that the closer the two entities are in animacy, the stronger the animacy bias will be. Research on similarity-based competition (Gennari et al., 2012) offers one possible explanation. This is the idea that competition will arise between two lemmas for lexical retrieval because of their common semantic features, leading one to be inhibited. When a speaker is required to talk about two entities that are semantically similar (e.g., man and woman), they will be more likely to separate them, for instance via the use of a passive structure (“the man who is being hit by the woman”) rather than a structure where the nouns are closer together such as in an active object relative clause (“the man who the woman is hitting”). Semantic similarity has also been linked to increased cognitive load for the speaker (however, see also Costa et al., 2005). While investigating word retrieval, Smith and Wheeldon (2004) found that, when describing a picture involving two objects (e.g.,
moving in some direction together), speakers tend to take longer to plan an utterance when doing so involves retrieving nouns for objects that are conceptually related (e.g., “the saw and the axe move down”) rather than for objects that are conceptually unrelated (e.g., “the saw and the cat move down”).

It is possible that the relative social similarity between human and humanoid may have created greater difficulties for production for describing these interactions than those between human and Lego, on account of the smaller animacy disparity. Indeed, post-hoc analyses of reaction times in Experiment 7 revealed that participants were slower to produce descriptions of human/humanoid interactions than human/Lego interactions ($M = 9338.93$ ms [SD = 3957.66] vs. $M = 9039.06$ ms [SD = 3952.63], $\beta = -303.01$, $SE = 83.16$, $df = 15876.96$, $t = -3.64$, $p < .001$). However, we did not find a significant difference between these types of trials in Experiment 6, so the effect of animacy disparity on production times, at least in this study, is inconclusive.\(^{13}\)

Alternatively, the effect of anthropomorphism could be explained by realistic expectations. It could be that participants perceived interactions between humans and humanoid robots as something that happens meaningfully in the real world, as these types of robots tend to be suggested as the new way forward for human/robot interactions, either through media or through actual use in everyday examples such as in healthcare, education and supermarkets (Broadbent, 2017). Even with limited understanding of social robots, it is likely that participants arrived with preconceived notions on what these types of machines can do. Lego, on the other hand, is arguably

\(^{13}\) This analysis was conducted using linear regressions. For Experiment 7, this was of the form Reaction Time $\sim 1 + \text{Anthropomorphism} \times \text{Narrative Condition} + (1 | \text{Subject}) + (1 | \text{Verb})$. The results also showed a significant intercept ($\beta = 9122.76$, $SE = 359.05$, $df = 97$, $t = 25.41$, $p < .001$), a non-significant effect of narrative condition and a non-significant interaction between anthropomorphism and narrative condition. For Experiment 6, this was of the form Reaction Time $\sim 1 + \text{Anthropomorphism} + (1 + \text{Anthropomorphism} | \text{Subject}) + (1 | \text{Verb})$. Results showed a significant intercept only ($\beta = 8355.6$, $SE = 352.7$, $df = 111.3$, $t = 23.688$, $p < .001$).
more associated with imagination and play. In this way, the capabilities and limitations of humanoid robots may be more fixed to the real world, whereas participants may have viewed the less-known capabilities of the Lego robots with more flexibility and, perhaps, with more willingness to suspend their belief when describing a human and a Lego robot interacting than a human and a humanoid.

Another possibility is that perceiving human/robot interactions has effects on production that are not comparable with the cases that are usually explored in relation to the animacy bias, for instance, descriptions of humans with animals or humans with other kinds of objects. What is particular to perceiving robots that does not apply to a bear stalking a hiker, or a dog pursuing a postman (Gleitman, 2008)? One consideration is the effect of discomfort that social robots can inspire. We consider that participants may have mentioned the Lego robots first more often than the humanoid robots, because interactions with the more anthropomorphic humanoid robots created a sense of eeriness in participants. When people encounter non-human entities with a closeness to humanness, such as artificial intelligence, machines and CGI, they may experience negative feelings such as discomfort or eeriness upon perceiving a clearly non-human property or behaviour (Mori, 1970). This might include an unnaturalness of artificial speech, robotic gestures in an android or a perceived soullessness or creepiness of CGI characters in film and TV – for example, with audience reactions to the 2019 film Cats with CGI human/cat hybrids (Cantrell & Hawkes, 2021). This mismatch between a person’s expectations of an object and its actual capabilities creates negative tension known as the uncanny valley, which has been further attributed to a process known as category ambiguity, or a difficulty in deciding how to categorise an entity (Yamada et al., 2013; Strait et al., 2017).
Recent research has found that human-like machine voices and appearances can cause eeriness for listeners associated with uncanny valley effects (Betriana et al., 2021; Clark et al., 2021). However, the uncanny valley has not yet been rigorously tested in relation to speakers, and so it is unclear how this type of unease might affect language production. We may consider that category ambiguity or uncanny valley effects further influenced participants to mention the human first over the robot. This reaction, according to the uncanny valley hypothesis, would have been more relevant for the humanoid robots rather than the Lego robots, thereby offering a potential explanation for why the humanoid robots were mentioned first less often than the Lego robots. We may also consider that the presentation of the humans in these interactions may have also created a sense of unease, particularly in Experiment 7 where they were described, not as human, but as human-like. The uncanny valley remains somewhat uncharted territory, so this explanation would benefit from further study.

4.4.2. Attitudes towards robots and perceptions of animacy

Next, we will discuss effects of individual attitudes and perceptions on mention order. In Experiment 6, this involved investigating effects of individuals’ attitudes towards robots on the human-first bias. We did not find strong evidence that a human-first bias could be predicted by a speaker’s attitudes towards human-robot interaction. None of the three NARS sub-scales were significant predictors of the human-first bias. Thus, our results do not suggest that people who have more favourable attitudes towards robots will be more likely view a robot as an agent and to describe interactions such as these from the robot’s perspective rather than the human’s.

In Experiment 7, our investigations involved testing perceptions of animacy of the human and robot groups as predictors of mention order. First, we found that relative
empathy (whether participants empathised more with the human or the robot group) was a significant predictor of mention order. The more someone empathised with the robot group over the human group, the less likely they were to mention the human first. This was apparent for both narrative conditions, although the effect of empathy appeared strongest on the robot-oriented condition, revealing a robot-first bias when participants showed big differences of empathy between the groups in favour of the robots (i.e., strong empathy towards the robots and weak empathy towards the human). This finding signifies empathy as a crucial factor in perceived animacy, as it entails not only the willingness to consider the mind of another entity, but to share in the feelings and emotions of that entity. In this way, we can also explain relative empathy in terms of conceptual accessibility, as a more accessible entity will have more predicates associated with it and thus more grounds for empathising than an entity that can do and experience fewer things.

Similarly, we found that participants who showed more dehumanisation towards the human group were less likely to mention the humans first. Thus, perceived humanness was a reliable predictor of the human-first bias. This was apparent across both narrative conditions, suggesting that manipulation is not required for dehumanisation to affect mention order. We also found that scores on the story traits measure was a stronger predictor than scores on the general traits measure. Interestingly, this corresponds with investigations of the narrative manipulation’s effectiveness. Participants in the robot-oriented condition were significantly more likely to dehumanise the human group than participants in the human-oriented condition, giving lower ratings for both the general traits and story traits subscales, but they were generally more likely to demonstrate dehumanisation on story traits than general traits.
This study cannot tell us which manipulated feature of animacy was responsible for interrupting the human-first bias. In the future, it may be possible to untangle these elements by emphasising only one element in the narrative. However, it is also likely that multiple properties are required to alter perceptions of animacy, or to sufficiently do so for effects to be observable. Additional work is required to determine the extent to which animacy can be attributed to one specific factor, or if it is the amalgamation of a range of properties – probably the more likely account.

Comparisons of animacy ratings within narrative conditions (comparing robot ratings with human ratings for each condition) may help explain the mention order patterns found in this experiment. Our data suggest that animacy ratings for humans, as displayed in Figure 9, were indicative of the mention order biases we found in the language task. While participants in the human-oriented condition perceived the humans as being much more animate than the robots, participants in the robot-oriented condition did not show a large difference between the groups for these animacy properties. From these ratings, we could have expected a high proportion of trials to show a human-first bias in the human-oriented condition, as the mean animacy score for humans from the human-oriented condition (averaging across agency, experience and empathy) was 72%. Likewise, we could have expected around half of trials to show a human-first bias in the robot-oriented condition, as the mean animacy score for humans from the robot-oriented condition was 53%. Indeed, we can argue that these scores are consistent with human-first proportions we obtained from the language task, which were 78% in the human-oriented condition and 57% in the robot-oriented condition.

Generally, participants in the robot-oriented condition rated the robots as having slightly more agency than the humans and empathised slightly more with the robots than the humans, but they rated the humans as having slightly more experience than the
This latter result might suggest that perceived experience in a non-human entity is harder to manipulate through narrative than other factors such as agency and empathy. Perhaps experience is so keenly associated with being human that our relatively simple short story was not convincing enough, therefore failing to encourage participants to consider the robots in the robot-oriented condition as having high experience and the humans as having low experience. Mappings of mind perception provided by Gray et al. (2007) support this notion; robots were rated as having middle-ground agency but no experience, and high experience was solely attributed to animals and people, in other words things that were alive.

While we may have been successful in reducing the human-first bias in Experiment 7, we did not outright reverse it. This implies that a stronger manipulation, addressing other properties, may be necessary to create the perception of robots that are more animate than humans or to influence a robot-first bias. For instance, we may consider alternative cues of similarity to the speaker (e.g., by showing robots as part of the same group as participants). We may also consider clearer emotion presentations, especially considering the faces of the robots in these trials were far less expressive than those of the humans. Emotion recognition is regarded as an important element of empathy and anthropomorphism, and an inability to do so has been attributed to negative perceptions of simulated humanness, such as audience reactions to computer-generated imagery, or CGI, (Cantrell & Hawkes, 2021). It is possible, then, that participants in the robot-oriented condition did not view the robots as much more animate than the humans, because they lacked expressive facial gestures.
4.4.3. Effects of valence, age, and a Like Me bias

We did not find strong evidence that valence affected mention order in this study. However, valence may still offer a promising area of research for future studies addressing descriptions of dyadic interactions, including those between humans and robots. For instance, we might consider the importance of the type of intentionality on perceptions of human-robot interactions, such as by comparing the intention to cause harm versus the intention to help or the intention to bring joy. Research into human-robot aggression has found that by increasing anthropomorphism of a robot via increasing properties of mind attribution, humans are more likely to show aggression towards the robot in the face of threat (Keijsers et al., 2021). What effect would we see on mention order if speakers were describing more explicitly aggressive interactions or, on the flip side, interactions of care? Effects of valence on mention order and human-robot interactions should be tested with a wider range of normed events to explore this further.

By and large, older participants were more likely to mention the human first than younger participants, but this was only statistically significant in Experiment 6. It is possible that the design of this study was not powerful enough to detect effects of age. For instance, the oldest participants tested were 73 in Experiment 6 and 74 in Experiment 4 – perhaps results would be different if testing participants over this age group. The distribution of ages within the sample may have also been a factor, considering the majority of participants were middle-aged. It is also possible that this mention order bias simply does not differ reliably across generations.

We did not find strong evidence that the human-first bias was modulated by a Like Me effect; participants mentioned the human first regardless of whether the human was a man, a woman, a Black person, or a white person. This is particularly interesting
considering our findings from Chapter 2. Our results in this study suggest that the Like Me gender and race biases will be eliminated when a speaker is describing an interaction between a human and a robot. Instead, it appears that the most meaningful categorisation a speaker is making in this instance is that of “human” and “not human”.

4.4.4. Effects of gender

Identity effects appeared elsewhere – in relation to the participants. Namely, this study revealed some unexpected effects of participant gender, whereby male participants appear to have been driving most of the animacy manipulation effects. First, in both experiments, we found a significant interaction between anthropomorphism and participant gender. Male participants showed a greater likelihood to demonstrate a human-first bias on human/humanoid trials than on human/Lego trials, and female participants showed little effect of anthropomorphism. We may consider that, in the norming study, the humanoid robots were rated as significantly more “female” and “feminine” than the Lego robots, who were instead rated as more “masculine” and “male”. Findings in Chapter 2 on the Like Me bias have revealed that female participants are more likely to mention a woman first when describing woman/man interactions, while male participants are more likely to mention the man first. We considered that mention order in the current experiment may have been influenced by gendered perceptions of the robots, with male participants more likely to mention the Lego robots first than the humanoid robots because they had a higher average rating of “male” and “masculine” than the humanoid robots.

We deconstructed this further during a post-hoc descriptive statistical analysis to determine if Pepper, the robot that received the highest ratings of “female” and “feminine” at pre-test, was also the robot who male participants were least likely to
mention first. This was not the case – male participants were least likely to mention NAO first, a humanoid robot with norming ratings more skewed towards maleness and masculinity, although less-so than the Lego robots (M = 0.73), followed by Pepper (M = 0.69) and then the two Lego robots Princess (M = 0.67) and Gresh (M = 0.68). The same applied to female participants (NAO M = 0.75, Pepper = 0.72, Gresh = 0.73, Princess = 0.73). Thus, perceived gender of the robots is unlikely to account for the finding that men appeared highly affected by anthropomorphism and the animacy disparity whereas women did not. Perhaps, instead, if the uncanny valley phenomenon is to be considered a possible explanation for the effect of anthropomorphism, we should also consider if this effect is gendered. So far, there is no evidence to indicate that this is the case, however that is due to a gap in this area of research rather than null results.

We also found effects of participant gender by way of interactions with scores from the post-study questionnaires. In Experiment 6, participant gender interacted with the NARS measure of attitudes towards emotional interactions with robots. More positive attitudes towards emotional interactions with robots contributed to a lower human-first bias in men but contributed to a higher human-first bias in women. In Experiment 7, we found a significant interaction between participant gender and the general traits measure from the dehumanisation sub-scale. Male participants who had shown greater dehumanisation were less likely to mention the human first, while female participants tended to show the opposite effect. For women, higher dehumanisation scores on the general traits subscale correlated with a greater likelihood of a human-first bias. This behaviour from female participants poses an incongruence with dehumanisation and mention order theories.
In Experiment 7, relative agency and experience were not predictors of mention order on their own, but they both interacted with participant gender. Male participants who showed lower perceived relative agency (i.e., indicating a smaller difference between humans and robots than with higher relative agency) were less likely to mention the human first, although this was only the case for the robot-oriented condition, not the human-oriented. Additionally, male participants who showed lower perceived relative experience were less likely to mention the human first for both narrative conditions. In fact, when male participants in the robot-oriented condition rated the robots are having much higher mind than the humans, they tended to mention the robot first.

Results from the female population were less clear: for both relative agency and experience, only those in the robot-oriented condition showed an effect of these ratings. This is understandable, considering the robot-oriented condition actively worked to change perspectives of agency and experience. Female participants in the robot-oriented condition who showed lower perceived relative agency were less likely to mention the human first, though to a lesser extent than male participants. Unexpectedly, if they showed lower perceived relative experience, or in fact rated the robots as having more experience than the humans, they were more likely to mention the human first. This is the opposite effect to what we had predicted, and while our analyses of the narrative manipulation indicated that we had not successfully encouraged participants in the robot-oriented condition to ascribe greater experience to the robots than the human, this seems like an unlikely explanation for this particular result considering the male population behaved as predicted.

Results from both experiments suggest that the men in the samples were driving the effects of anthropomorphism, relative agency, relative experience and
dehumanisation. In Experiment 7, we found that men were much more susceptible to the narrative manipulation than women regarding their mention order biases. As with the emotional interactions finding, men and women often showed opposite patterns of behaviour to each other for each of these effects; men were less likely to show a human-first bias the more they attributed mind to the robots, but such mind attribution often fortified the human-first bias of women.

Why might men have been more affected by manipulations of anthropomorphism and animacy than women? It is hard to say from our data, however we might speculate that these results support the assertion that person perception activates a different neural network than processing non-human human-like entities. To some extent, our results are inconsistent with previous research that has found women to be better at mentalising than men, as determined by a better ability to empathise (Abu-Akel & Bo, 2013; Baron-Cohen & Wheelwright, 2004). However, women in the robot-oriented condition were not more likely than men to mention the robot first, despite having been cued to perceive the robots as more animate and deserving of empathy. The gender differences we found are interesting but difficult to untangle, however they do suggest that normal prediction of mentalising for women versus men are not attributable to perceptions of robots as social agents.

4.4.5. Memory for names

Lastly, our control analyses indicated the importance of considering memory effects on mention order. This analysis was as important in the current study as it was in our previous studies, because a possible explanation for mention order biases could be that the first name mentioned was the easiest name to retrieve. Our results in Chapter 2 on the Like Me bias do not appear to suggest that mention order was driven by a
memory for names, as participants did not tend to show a better ability at remembering names of people of their own gender and race, despite mentioning them first. In the current study, participants in Experiment 6 were significantly better at remembering the names of the humans, who had common English names (e.g., *Fred*), than of the robots, who had fictitious robot-related names (e.g., *Cybe*). In Experiment 7, in which we assigned both humans and robots fictitious names, the difference in memory accuracy for human and robot names was much smaller compared to Experiment 6, highlighting the importance of standardisation the names of human and robot figures. However, participants did still show a better ability for remembering the human names than the robot names. Narrative condition did not appear to affect this accuracy, therefore memory for names was not affected by how prominent those names had been in the narrative (i.e., the names for the story protagonists – humans in the human-oriented condition and robots in the robot-oriented condition – appeared more frequently than the names for the other group).

That participants showed a better ability for remembering the names of the humans over the robot, regardless of name familiarity or exposure through the narrative, is consistent with the finding that there are distinct differences between memory for people and memory for non-social information (Wyer, 2014). Although the robots were presented as social agents, the difference between human and non-human appeared to affect participants’ ability to remember information about both groups. We might also consider that perceptual factors had an effect on memory accuracy. It is possible that participants found it easier to remember the humans’ names, because the humans were easier to distinguish apart than the robots. There were similarities in aesthetics between the humanoid robots (both white in colour and metallic) and between the Lego robots (both colourful and plastic-looking), and there would have been a level of unfamiliarity
for participants regarding the robot figures which could have made learning associations between them and their names more difficult. Thus, we cannot rule out that participants may have been more likely to mention the human first than the robot because they had remembered the human names better, however results from Chapter 2 suggest that conceptual accessibility was likely the greater determinant of mention order, rather than memory effects.

4.4.6. Summary

The current study provides important insight for how conceptual accessibility influences language production, first by showing that the animacy bias applies for descriptions of human/robot interactions. Our results have also demonstrated that the gap in animacy between two entities, or the animacy disparity, acts as a further predictor of mention order biases. Crucially, our results also suggest that considering someone as more similar to ourselves in mind and ability can also influence us to think of that person as more capable of doing certain things than others who are less socially similar. Moreover, we have identified the flexible nature of inherent accessibility and animacy. A short story narrative allowed us to alter standard perceptions of animacy through anthropomorphism of the robot group and dehumanisation of the human group, thereby reducing the human-first bias. However, our results also indicated that some animacy properties were easier to manipulate than others, with participants not fully buying into the idea that the robot protagonists would have more experience than the human clones.

Understanding animacy will benefit not only psycholinguistic research, but also that of social psychology, person perception and social robotics. In doing so, we can continue to explore how different people (or robots) are perceived on different hierarchies of mind, and the consequences of these distinctions. The theory of
conceptual accessibility demonstrates that first mention is tightly bound with what a speaker considers plausible about an entity’s behaviour or abilities. Thus, who a speaker chooses to mention first may reveal just as much about the speaker as it does about the described.
5. General Discussion

When describing an interaction between two people, a speaker must rapidly decide how to conceptualise and represent that event. For instance, they may have to choose who is the sentence subject and direct object (e.g., “the student talked to the professor”) or if both people are attributed subject status (e.g., “the student and the professor talked”). Either way, the speaker must decide who to mention first and second. Mention order both determines where attention is initially directed (MacWhinney, 1977) and can also signify from whose perspective the interaction is described. Previous work from a range of domains, including psycholinguistics and social psychology, has identified how conceptual factors such as prototypicality, imageability, relevance and animacy affect mention order (e.g., Iliev & Smirnova, 2016; Bock & Warren, 1985; Onishi et al., 2008; Kesebir, 2017; Hegarty et al., 2011; Prat-Sala & Branigan, 2000; Branigan et al., 2008; McDonald et al., 1993). This thesis aimed to unify these domains in order to investigate social biases of mention order, addressing gender, race and human animacy through seven experiments. My results provide novel evidence for a new conceptual factor that can influence who speakers talk about first when describing a dyadic interaction: relative social similarity between the speaker and the people who feature in their description. They also show that mention order reliably influences semantic inferences, with comprehenders making judgements of agency from first mention. This thesis demonstrates that language production and comprehension is fundamentally influenced by hierarchies that exist within society and by how people identify with social groups.

In Section 5.1, I will summarise the findings from the three empirical chapters and their respective studies. First, I will focus on results from the language production studies and the evidence they give for a Like Me bias of mention order. Second, I will
focus on the results from the comprehension study and how this shows that mention order influences judgements of agency. Then, in Section 5.2, I will discuss explanations of the Like Me bias and first-mentioned agency bias, including considerations of how our findings contribute to existing language theories. I end on directions for future research and acknowledge how this novel work equally contains limitations and provides important methodological contributions for psycholinguistic research.

5.1. Summary of empirical findings

5.1.1. Effects of social cognition on mention order

In our language production experiments (Chapter 2 Experiments 1, 2 and 4, and Chapter 4 Experiments 6 and 7), we used a picture description task to measure mention order biases. These experiments tested social hierarchies and relative social similarity as predictors of mention order, addressing gender and race in Chapter 2 and human animacy in Chapter 4.

Participants in Chapter 2, who were Black and white women and men, used symmetrical predicates to describe images depicting two people interacting, who either differed in gender (woman/man) or in race (Black/white). On gender trials, we determined relative similarity as participants being the same gender as one person and a different gender to the other (e.g., a female participant viewing an interaction between a woman and a man). On race trials, we determined relative similarity as participants being the same race as one person and a different race to the other (e.g., a Black participant viewing an interaction between a Black and a white person). We predicted that speakers would show biases in their mention order that are driven by facets of
social cognition, including those that reflect conventional social hierarchies (a man-first bias on gender trials and a white-first bias on race trials), or a Like Me bias to mention the person most similar to themselves in social identity first (a Like Me gender bias on gender trials and a Like Me race bias on race trials).

Importantly, these factors are not mutually exclusive. Our results provided some support for the first hypothesis but more extensive support for the second. Although we found an effect of a gender hierarchy on mention order, represented by a man-first bias, we did not find the same effect for a racial hierarchy, which would have been represented by a white-first bias, suggesting that conventional social hierarchies in general are not a reliable predictor of mention order. Furthermore, on gender trials, although speakers tended to show a man-first bias in their descriptions, this was mostly driven by male participants and was largely moderated by the Like Me gender bias; speakers showed a tendency to mention the person of their own gender first and the person of the other gender second. On race trials, speakers’ descriptions showed a Like Me race bias; they tended to mention the person of their own race first and the person of the other race second.

Our results also prompted us to test effects of social group on name recall during the memory task, to determine if participants were better at naming one group over another. Specifically, we tested if they were better at remembering the names for women or men, or Black people or white people, or if participants were better at remembering the names of the people of their own gender or race over the names of the people from their outgroups. Interestingly, we did not find evidence that relative social similarity influenced memory accuracy. However, results suggested that participants tended to be better at remembering names for men rather than women (this effect was marginal in Experiment 1 and significant in Experiment 2 and 4).
Our main findings show consistencies with previous findings in some ways, but inconsistencies in others. Notably, as similar to Iliev and Smirnova (2016), we found a man-first bias in mention order on gender trials. However, comparisons of our results indicate big differences in the frequency of this bias for transitive and conjoined sentences (our studies) compared to binomials (Iliev and Smirnova’s study). In their study, participants answered the question “what are the two genders?” by mentioning the male gender before the female gender 91% of the time. In our study, a man-first mention order, although significant, occurred on only 54% of trials in Experiment 1 and on only 52% of trials in Experiment 2. In Experiment 4, the man-first bias was not significant, occurring on 51% of trials. This suggests that the huge skew towards naming male lexical items first in Iliev and Smirnova’s study was likely driven by frozen word orders that see “man and woman” and “male and female” occurring more frequently in English than “woman and man” and “female and male” (Cooper & Ross, 1975; Mollin, 2012). The results of our study also show that a man-first bias can appear without the influence of frozen orders, since participants used proper nouns for the individuals they were mentioning, rather than common nouns of lexical gender, and did so without conjunctions.

Importantly, in both our study and theirs, results indicated that women were significantly less likely to demonstrate a man-first bias than men in production. And, in fact, Iliev and Smirnova’s study showed that, across a range of categories (gender, religion, university affiliation, political alignment, brand preferences), people demonstrate a first mention bias for nouns that are relatively more associated with themselves than those that are less associated (e.g., Christians on Christian websites using the order “Christian and Muslim” far more often than the opposite direction). Both their study and ours demonstrate that specific characteristics of the speaker can
affect their mention order. Our work extends these previous findings, by revealing that speakers demonstrate a first mention bias, not only for categories and groups that are more related to themselves, but also for individuals who are more socially similar to the speaker.

Additionally, the man-first bias demonstrated by the participants in our study is consistent with arguments Hegarty et al. (2011) have made from their research into gender effects on mention order. They also found that a man-first bias, which was modulated by social conventions. Participants named couples who they were told were either traditional (“conform strictly to gender scripts about how the two genders should behave”) or non-traditional (“deviate radically from gender scripts about how the two genders should behave”). They found that participants showed a tendency to mention the man first for traditional couples but not non-traditional couples, and that male participants were more likely to do so than female participants. However, they did not include a control condition, for instance giving information about the couples that was irrelevant to gender roles, and so we can only draw conclusions from this experiment about how a man-first bias manifests according to perceptions of traditional gender roles, and not more generally. In our study, we presented dyads who were as similar as possible in all other respects, from height to clothing and facial expressions, thereby demonstrating that gender has an effect on mention order without additional lexical background information.

Hegarty et al. (2011) have also proposed that, when familiarity is a factor, people will be more likely to mention a person of their own gender rather than showing a general man-first effect. They collected Christmas cards from participants, testing how people named heterosexual couples they knew personally. Results showed that people were more likely to mention the person they were closer to in relationship first, and that
this correlated with an own-gender effect (i.e., women tended to address Christmas cards to couples mentioning the woman first and the man second, and men did the opposite). They then tested this effect in an experimental context, asking participants to write down the names of heterosexual couples from their friendship and familiar circles, and imaginary heterosexual couples. Their results showed that participants often mentioned the person of their own gender first when they were naming couples who were known to them, but this own-gender effect did not appear when they were naming imaginary couples. However, our findings show that knowing someone personally and having relational closeness is not required for a speaker to mention a person of their own gender first. Instead, our evidence suggests that the Like Me gender (and race) bias participants demonstrated in our studies is more likely the product of group identification, rather than familiarity effects as proposed by Hegarty and colleagues.

Crucially, we were able to replicate the Like Me gender and race biases we found, uncovering the effect of relative social similarity in three separate experiments, including using different visual stimuli and slightly different verbs and actions. As a result, we are confident that these results were modulated by our controlled factor of relative similarity rather than confounding linguistic effects, such as frozen orders in binomials or name familiarity, contextual effects such as the specificity of naming couples, or a Type I error. Thus, our results offer a generalisable trend of mention order for describing social interactions.

5.1.2. Effects of valence

Overall, we did not get very strong effects of valence across these experiments. Our investigations began as exploratory; the events were not pre-tested, and so our three categories of valence may not accurately represent intimate, casual and negative events.
Nevertheless, some results were consistent with previous work, in that they suggest there may be something specific about a man-first bias for name orders in a romantic relationship context. Although participants were at no point told that the pairs of individuals were in romantic couples, the trials coded as “intimate” (hug, kiss, dance and marry in Experiment 1, flirt, kiss, dance and marry in Experiment 2, and hug, kiss and seduce in Experiment 4) consequently framed the interacting pairs as resembling romantic couples. Interestingly, in all three experiments, we found that the man-first bias occurred more often for intimate events than casual or negative events. In Experiment 1, a significant interaction between syntax and valence showed that this was the case only for transitive sentences, not conjoined sentences. In Experiment 2, this interaction was not significant; a man-first bias was most likely for intimate events than the other events, for both transitive and conjoined sentences. Finally, in Experiment 4, there was again a significant interaction between syntax and valence. A man-first bias was more likely for intimate events than casual or negative events for active sentences (e.g., Fred is kissing Kate), but not for passive sentences. In fact, for passive sentences, a man-first bias was more likely for negative events than intimate or casual events (e.g., Fred was being shouted at by Kate).

These findings suggest that when a man-first bias does emerge in mention order, the driving factor may be the distinction of romantic couple over platonic pairings. This is consistent with Hegarty et al. (2011), who showed that women and men demonstrated a man-first bias when naming unfamiliar couples. They argue that when a speaker does not know the couple personally, they are more likely to be influenced by a conventional gender hierarchy in their mention order choices. Perhaps this is why in our study, in which participants were also naming unfamiliar people, the man-first bias tended to occur most often on intimate trials.
Our results also build on the findings of Oeberst and Matschke (2017), who showed that when people name wars in which their ingroup was involved, they are more likely to mention their ingroup first and the outgroup second. Importantly, however, our investigation of the Like Me bias shows that mention order, as influenced by group identity, need not be driven by stakes as high as conflict. Through testing events of different valences, importantly including some comparatively low-stakes and casual events such as talking and meeting, we have demonstrated that the relative social similarity effect occurs beyond representations of conflict.

War is an emotive and affective concept, where “sides” and dichotomous perceptions (i.e., winner and losers, and those in the right versus those in the wrong) are central to how we think about the conflicts. Thus, it is reasonable that ingroup identity might affect the dichotomy of first and second mention when referring to war; perhaps people like to view their own country historically as winners of wars, perhaps they wish to draw attention to their country as important. Numerous factors might explain why someone would mention their own country first in naming conflict. It is possible, then, that Oeberst and Matschke’s findings are less about conflict per se and are more about the accessibility of an ingroup over an outgroup. This could be a case of linguistic accessibility, in that people may be likely to hear or think about the name of their own country more frequently than the name of another country. Findings from their minimal group paradigm, in which participants were assigned to one of two fictional genetic subtypes, were used as evidence for a Me First effect on conflict naming. When asked to name a conflict between the two groups, participants tended to mention their fictitious ingroup first. However, these findings do not rule out the possibility that accessibility modulated mention order biases, as the participants’ ingroup was made more salient through the task, when participants were asked to think about their ingroup.
5.1.3.  *Mention order and syntactic form*

Vitally, this thesis builds on what is known about biases in production by addressing how social factors affect mention order biases in grammatically complex, full sentences. Not only, then, do people mention someone who shares a group identity with them first when producing binomials such as noun pairs (Iliev & Smirnova, 2016), couple names (Hegarty et al., 2011) or hyphenated conflict names (Oeberst & Matschke, 2017), but they do so when they are interpreting visual events and distinguishing between the roles of sentential subject and direct object, and of thematic agent and patient, using verbs and syntax.

We tested the Like Me gender and race biases on different syntax structures, investigating how they would manifest for transitive sentences (“Kate is hugging Beth”), conjoined sentences (“Kate and Beth are hugging”) and passive sentences (“Kate was hugged by Beth”), and found that speakers demonstrated this bias across syntactic forms. Consequently, our results indicate that speakers will show a tendency to mention a person who is relatively more similar to themselves in social group first over a person who is less similar, regardless of whether there are distinctions of grammatical role or thematic role between the two people named.

The fact that syntactic form did not reliably modulate the Like Me bias in first mention is interesting, in light of previous studies that sometimes do find syntactic influences on mention order biases, and sometimes do not. For instance, Bock and Warren (1985) only found effects of imageability on mention order when the two differentially imageable nouns differed in grammatical function, such as for transitive sentences, and did not find such effects in constructions like conjoined noun phrases where both nouns had the same role. On the other hand, Onishi and colleagues (2008) found effects of biased mention order, as it related to prototypicality, only on conjoined
noun phrases and not when the nouns differed in grammatical function. Interestingly, however, there is evidence that animacy influences mention order regardless of syntactic form – Prat-Sala and Branigan (2000) and Tanaka and colleagues (2011) have shown that speakers maintain an animacy bias in both active and passive sentences, which is consistent with our findings.

Relative social similarity and animacy are arguably conceptually far more similar than prototypicality and imageability, which may also explain these differences in syntax effects. Support for the parallels between relative similarity and animacy came from Chapter 4, which tested the Like Me bias using human animacy as a property of relative similarity. These studies showed that animacy modulates perceptions of relative similarity in a comparable way to constructs such as gender and race, through distinctions of “like me” as humans and “not like me” as non-humans. Thus, since the Like Me gender and race biases occurred on both active and passive sentences, as did the animacy bias found by Prat-Sala and Branigan, this may imply that the Like Me bias and the animacy bias have effects on production that are related, and indeed more comparable than other conceptual factors such as prototypicality and concreteness.

Crucially, since participants showed a Like Me bias across syntactic form, our results suggest that it is the sequence of first and second mention, rather than the assignment of grammatical roles, that is influenced by relative social similarity. By mentioning the person who is the same gender or the same race as they are, the speaker directs the focus of the sentence towards themselves.

5.1.4. Visual cues of agency

In Experiment 3, we considered a possible cognitive explanation for the mention order biases found in previous experiments; that participants perceived certain people as
more agentive than others on view alone and were consequently prompted to mention these people first. However, our results suggest that the Like Me bias is not driven by visual cues of agency. To test this, we used the same images of women and men, and Black people and white people interacting as in Experiments 2 and measured whether participants would be more likely to perceive a man as the more likely causative and volitional agent than a woman, and a white person as more agentive than a Black person (as per our earlier hypothesis of fixed social hierarchies). Alternatively, we considered that they might show a Like Me bias to perceive a person of their own gender, or of their own race, as the more likely agent than a person of a different gender or race. Our results did not provide strong support for either hypothesis, or either feature of social identity, and so it seems unlikely that speakers showed a Like Me gender and race bias due to perceptions of agency from the visual stimuli. This does not necessarily mean that visual cues of gender and race do not affect perceptions of agency, just that it is more likely that other effects were responsible for the Like Me bias.

5.1.5. Effects of animacy on mention order

In Chapter 4, we tested animacy effects on mention order with descriptions of human/robot interactions. Participants described images depicting a person and a robot interacting, also using symmetrical predicates. On half the trials, the robot was a more anthropomorphic humanoid and on the other half, it was a less anthropomorphic Lego robot.

As per predictions afforded by the animacy bias, participants showed a robust human-first bias overall. Using principles from the conceptual accessibility or animacy hierarchy, we also expected that the human-first bias would be stronger on trials with a larger animacy disparity, so those depicting a human interacting with a Lego robot,
rather than a human interacting with a humanoid. However, we found the opposite effect; the human-first bias was strongest on human/humanoid trials. This particular finding is important, because it suggests that the hierarchical explanation of the animacy bias may be more complex than previously understood. While it is highly predictable that a speaker will mention an entity higher on the conceptual accessibility hierarchy first rather than an entity further down, previous studies had not compared how relative placement along the hierarchy affects mention order. Our results suggest that the simple prediction – the more distinct two things are on the animacy hierarchy, then the greater the difference in mention order will be – is not quite right. In light of our unexpected findings, I propose that animacy disparity effects should receive more attention and further study.

In Experiment 7, we aimed to determine which elements of animacy are responsible for the human-first bias found in Experiment 6. Participants completed the same human-robot task as in the previous experiment, but beforehand they were exposed to a narrative manipulation that took one of two forms. Those in the human-oriented condition read a short story in which a human group was described as having more human animacy than a robot group, and therefore as more similar to participants. Those in the robot-oriented condition read a similar short story, except this time the robot group were described as having more human animacy than the human group, and were therefore presented as more similar participants. Cues of human animacy were primed from details in the stories, demonstrating the more animate group’s agency (i.e., their ability to plan and act) and experience (i.e., their ability to think and feel), using properties described by Gray et al.’s (2007) dimensions of mind perception, such as details of morality, thought, personality and consciousness. Then, all participants
completed the same task as in Experiment 6, using the same pictures that had previously elicited a robust human-first bias.

Priming participants to think of the robots as more animate than the humans caused a significant decrease in the human-first bias, although it did not remove it completely. This suggests that the human-first bias is driven by a range of different animacy features that include, but are not limited to, perceptions of mind abilities such as agency and experience, but also physical characteristics like appearance.

5.1.6. Effects of mention order on perceptions of agency

Having established that biases exist in mention order, in Chapter 2 we investigated if mention order has semantic effects on the comprehender (Experiment 5), testing if first mention influences judgements of agency. To do so, we used a rating scale to measure perceptions of agency according to two properties: causation and volition (Dowty, 1991), testing the extent to which first mention influenced these perceptions. We used the same names (e.g., Beth, Kate, Fred) and verbs as in Experiment 2 (i.e., symmetrical verbs such as hugging, arguing, playing) in transitive and conjoined sentences (e.g., Beth is hugging Kate; Beth and Kate are hugging) and measured if participants would show a tendency to rate the first-mentioned name as the more likely agent than the second-mentioned name. We also tested if syntax would modulate perceptions of agency, comparing transitive trials where there were clear thematic agent/patient roles and conjoined trials where these distinctions were not present. Our results showed that first-mention did affect perceptions of agency, with participants showing a tendency to rate this person as the more likely causational and volitional agent. The effect of first mention was also strongest on transitive trials, however we importantly found this effect on both types of syntax.
Importantly, since our language production experiments showed that first mention can be influenced by social factors, such as a man-first bias and a Like Me bias, it was necessary to test how social factors may also contribute to perceptions of agency from first mention. If certain people tend to be mentioned first, how does this affect judgements of important properties such as will, intention and accountability where language contains such biases?

Thus, we also tested effects of gender in Experiment 5, by showing participants sentences in which one name was for a woman and the other was for a man (e.g., “Kate is playing with Fred”). On these trials, participants were more likely to perceive the first-mentioned person as the more likely agent when they were a man rather than a woman. Importantly, both male and female participants were more likely to rate the first-mentioned person as agentive if they were a man. This result suggests that the gendered agency effect is more likely driven by the conventional social hierarchy of gender that presumes men to have more agency than women, rather than a Like Me gender bias whereby a person’s perceptions of agency are biased towards people of the same gender as themselves. Consequently, there is a disparity between comprehension and production in this respect. Where a speaker is more likely to mention their gender in-group first, they are not likely to perceive their gender in-group as the agent of the event unless they are male. It would be important to test semantic inferences in the context of race, too, to determine if agency judgements would follow a white-agent bias, or a Like Me race bias. The evidence we have provided supports the former prediction rather than the latter, but the results from our production studies also demonstrate how gender and race effects manifest differently in language, so both hypotheses should be considered.
Syntax also modulated this effect to some extent. With ratings of causation (e.g., “Fred asked to play with Kate”), we found this gendered effect of first-mentioned agency judgements for both transitive and conjoined sentences. However, with ratings of volition (e.g., “Fred asked to play with Kate”), participants were only more likely to rate the first-mentioned person as the likely agent when this was a man rather than a woman, if they were inferring agency from transitive sentences (e.g., “Fred is playing with Kate”) and not from conjoined sentences (“Fred and Kate are playing with each other”). This result suggests that people are less likely to infer volition from a conjoined sentence, compared to a description with distinctions of grammatical and thematic role.

Our results from this experiment show that people make reliable inferences from mention order alone, when judging who is more likely to be the agent of an event. These judgements will also be influenced by whether the first-mentioned person is a man or a woman, and whether they are judging if a person caused or wanted that event to happen.

To summarise, the seven experiments conducted in this thesis reveal important new information about the ways in which language may be affected by social constructs, including conventional social hierarchies and perceptions of relative similarity with the speaker. Next, I will discuss possible explanations for these biases and implications for both language research and wider society.
5.2. **Explanations and implications of this research**

5.2.1. **Explanations for The Like Me bias**

Why did speakers consistently mention socially similar people first? One possibility is that this is a presentation effect. Individuals might want to highlight the importance of their own group by giving ingroup members prominent sentence positions. Another, not mutually exclusive, possibility is that this is a matter of accessibility. A bottom-up explanation for the Like Me gender bias could be due to frequency. It is possible that women more frequently hear names for women than names for men, and men more frequently hear names for men than names for women, and so are more likely to mention a name they are exposed to more frequently. We might also consider that the names used in this study were all monosyllabic, and that this is relatively atypical for women’s names compared to men’s names (Crystal, 1993; Cutler et al., 1990), which might make the former more difficult to remember than the latter. It is possible, then, that a reoccurring man-first bias is partially attributable to the typicality of the names used in these studies.

It could also be that, in line with linguistic accessibility, participants initially mentioned the names of the people they remembered better, and that their memory for names was influenced by shared identity. For instance, Black participants mentioned the Black person first over the white person because they had stronger associations between the names of the Black people and their appearances on account of seeing that they were the same race. However, we did not find evidence to suggest that participants were better at remembering the names of the people of their own gender or race over the names of the people from their outgroups. Certainly, memory effects may have manifested differently in the memory task compared to the language task, but our
results do not convincingly point towards the explanation of differences in ease of name retrieval.

Instead, we must consider the role of conceptual accessibility. Having drawn parallels between the animacy bias and the Like Me bias, we may be justified to question the extent to which the Like Me bias is an effect of conceptual accessibility. So far, two other theories offer explanations for the Like Me bias. The first is that of the ‘Me First’ effect first proposed by Cooper and Ross, which states that nouns that are more central to the speaker will appear first (Cooper & Ross, 1975), and has since been attributed to gender biases in conjunctions (Iliev & Smirnova, 2016), couple naming (Hegarty et al., 2011) and conflict naming (Oeberst & Matschke, 2017). This theory has so far only been tested on binomials, but even so, the results from those studies suggest that the speaker’s own characteristics affect mention order, without distinctions of grammatical function. Then, there is the Like Me hypothesis in social cognition, which states that seeking out self-other equivalences are important for determining shared cognitive abilities (Meltzoff, 2007). Perhaps it is the perception of social similarity that makes someone a better starting point for speakers to build their utterances on, as in lexical incrementality (Gleitman et al., 2007), or that self-other equivalences affect how the relationship between two individuals is construed, which is required for structural incrementality (Griffin & Bock, 2000). We propose conceptual accessibility as another potential or joint explanation.

Current evidence might not necessarily suggest that social factors such as gender and race are determinants of conceptual accessibility, and we have not discussed them as such in this thesis, but Bock and her colleagues’ work do provide a lens through which to think about socially driven mention order biases: that observable and conceptual differences often exert effects on word order. We might consider that
speakers are more likely to mention people from their ingroup first, because they are better at imagining the abilities of people who are more like themselves over people who are more different. If so, this would link with the social cognition Like Me hypothesis as perceptions of shared cognitive abilities could be determinants of relative similarity.

As Branigan and colleagues have noted, animacy is an example of inherent conceptual accessibility, for which invariability across contexts and thus inflexibility is an attribute (Branigan et al., 2008). As we generally think about the world and the placement of the beings and objects within it, this fixed perception of animacy and a static animacy hierarchy is often true. However, Experiment 7 demonstrates how inherent accessibility can also be redefined, by altering the qualities of an entity and therefore what it can do. Previous studies have shown that conceptual accessibility can be changed through narrative, however these focused on altering derived accessibility, such as Prat-Sala and Branigan (2000) who manipulated discourse salience and prominence in text. In our experiment, however, we reconstructed the inherent accessibility of the humans and robots, showing that beliefs can reposition a typically non-social being as highly animate and a person as less animate.

Studies of dehumanisation further demonstrate the flexibility of human animacy, through social stereotyping by which perceptions of certain groups are considered animal- or machine-like, despite their category of human (Haslam, 2006). Our findings from Experiment 7 suggest that dehumanisation can affect a person’s inherent conceptual accessibility. An individual’s perception of animacy and their own beliefs about the abilities, experiences and intentions of an entity, or person, will be central to how they perceive the predictability of this entity or person and thus their conceptual accessibility. We could think about this as the less an individual or group is believed to
be capable of, the fewer predicates will be associated with it. This may in turn affect a speaker’s mention order choices, as people with lower predicable will, in theory, have lower conceptual accessibility. In this way, conceptual accessibility could be explored further to help understand how relative social similarity affects language production, and to identify social biases in the ways in which groups of people from different social groups are talked about by people from different social groups.

Exploring the role of conceptual accessibility on the Like Me bias would require a consideration of incrementality as the primary mechanism behind language biases. Our results from Experiment 3 do not suggest that participants mentioned the person who was more socially similar to themselves first because they perceived this person as the more likely agent. Nor do our results suggest that participants were more likely to mention these people first because their names were more linguistically accessible, as demonstrated through our memory task analyses. Instead, I propose that participants found the names of the people more socially similar to themselves easier to retrieve, because these people were more conceptually accessible to the speaker than those who were less similar. This ease of retrieval would manifest across incremental strategies; both structural incrementality (in which a speaker constructs their utterance after conceptualising the event and thus the relation between the entities) and lexical incrementality (in which the speaker’s starting point is influenced by perceptual factors such as where they looked first), as both are affected by the relative ease of retrieval for the entities in question. Consequently, I propose that the Like Me effect is a stable bias in production, because it will arise regardless of which incremental strategy the speaker uses. Importantly, our results suggest that applying psycholinguistic theories to the study of socially biased language will allow us to understand those biases better. We
need to consider dimensions beyond frozen conjuncts and Cooper and Ross’s (1975) Me First principle, if we are to understand how these patterns arise in the first place.

What does the Like Me bias tell us about social cognition? One possibility is that this bias is the result of ingroup favouritism (Maass et al., 1989; Maass & Russo, 2003). Speakers might mention someone from their in-group first because they are more likely to favour the perspective of someone who is more socially similar to themselves over someone who is less similar. However, this would not fully explain why the Like Me bias occurred on negative trials like arguing, fighting, and shouting. If ingroup favouritism was at play, we may expect that speakers would mention the person from their own group second on these trials for them to appear as less agentive and therefore less culpable for the aggressive interaction. However, our evidence did not point towards this.

Instead, I propose that the Like Me bias could be a humanising effect. In Experiment 7, participants who had dehumanised the human group more, regardless of whether they had been involved in the human-oriented or robot-oriented condition, were less likely to mention the human first. We also found that empathy was a significant predictor of mention order, with results suggesting that who participants mentioned first tended to reflect who they empathised with more – human or robot. These findings indicate that when a speaker perceives someone as being less human and therefore less like themselves, they will be less likely to mention them first.

Our work on the Like Me bias offers a framework for fields that seek to investigate descriptions of interactions, such as discourse analysis and psycholinguistics. Who is talking, and who is being talked about first, are two questions that could help reveal subtle social biases.
5.2.2. **Explanations for the first-mentioned agency bias**

Why do people perceive a first-mentioned name as more agentive than a second-mentioned name? This may be due to inferences the comprehender makes about the messenger’s intentions, perhaps consciously considering why the speaker mentioned one entity first and the other second or being subtly influenced by the order of mention. Considering the Gricean maxim of relation, whereby a speaker should attempt to say things that are relevant to the discussion (Grice, 1975), a listener may use this pragmatic rule to reasonably assume that the first entity mentioned by a speaker is more relevant than the second-mentioned entity.

Language experience may also play a role. Evidence for this comes from the finding that people tend to assume first-mentioned entities as more animate than second-mentioned entities (Johnson, 1967), and so there are likely to be a number of other properties associated with first mention. Additionally, conceptual accessibility means that more accessible entities (i.e., having qualities that are associated with agency) tend to be mentioned first, so people have a great deal of experience of hearing agents mentioned before patients, and may be driven by this statistical experience. Indeed, as Dowty (1991) proposed and Kako (2006) demonstrated, sentence subjects are more often associated with properties of prototypical agents (e.g., causes change), while direct objects are more often associated with properties of prototypical patients (e.g., is changed). People learn to link sentence subjects with thematic agents through linguistic experience (Foley & Van Valin, 2009; Bowerman, 1990). Thus, the participants in this study were already primed to consider the subject of the transitive sentences as agentive.

How does this apply to the same biased judgements of agency occurring for conjoined sentences, in which the two mentioned people did not differ in grammatical function? This result suggests that first mention not only sets the focus of a sentence but
also has semantic effects on the comprehender that are independent of grammatical function (c.f. Bock & Warren, 1985). I argue that the starting point of a sentence (MacWhinney, 1977) sets the focus for a comprehender and that it is this initial focus that drives how an event is construed.

This contrasts with prior work from comprehension studies, which argue that the second-mentioned entity tends to receive the privileged position of sentence “focus” for a perceiver encountering an utterance. This is due to the tendency for new information to appear towards the end of a sentence in English, for example in the case of “Mary likes Sue”, where Sue is considered the more intuitively interesting piece of information (Jackendoff, 1972; Krifka, 1992; Rooth, 1992; Carlson et al., 2009). What distinguishes our work from this, however, is that our emphasis is specifically on role judgments, rather than the perception of new information. Sue might be the relevant figure in the question of who is liked, but Mary is undoubtably the agent of the sentence. In asking the participants in Experiment 5 to rate who they thought caused the interaction or wanted it to happen, their judgement, and attention, was guided by first mention.

The finding that judgements of agency are influenced by first mention, regardless of the existence of grammatical function distinctions, has potential methodological implications for research on social biases in semantics. Agency is one of many possible assumptions a listener or reader might make from first mention. It is an umbrella term, which could comprise of anything from intentionality and ability to fault and culpability. Even an innocuous utterance such as “Fred and Beth are meeting” can create perceptions of power and objective, and this becomes more serious when we talk about harmful or violent events using symmetrical verbs where agency is somewhat ambiguous, such as “Fred and Beth are fighting” or two groups described as “clashing”. When the utterance does not explicitly make it clear who the instigators of violence are,
and who may be in a position of defence, the comprehender may make their own inferences from the order of mention. For researchers committed to studying social biases in semantics, our work highlights the importance of measuring perceptions of agency from mention order, as it is a simple but effective way of creating impressions.

5.2.3. Future directions

There are numerous directions that could be taken to extend our findings and further scientific knowledge. First, there is considerable scope for further characterising the Like Me bias, as there are a multitude of ways in which a speaker can be more socially similar to one person and less socially similar to another, not just gender and race. The Like Me gender bias can also be expanded to investigate different presentations of gender – how might a transgender woman or man, or a non-binary person respond to these tasks, for example? Equally, the Like Me race bias could be tested on a range of different racial groups, ethnicities and nationalities, for instance with mixed race participants, or by comparing people who have grown up and live in populations where their ethnic identity is the majority group versus people who have consistently been the minority in a population.

Investigations of the Like Me bias and human animacy need not be restricted to human and non-human entities, as demonstrated by the literature on dehumanisation (Haslam, 2006). What mention order biases would manifest when speakers describe interactions between groups that have and have not been typically dehumanised in society, such as people who have homes versus people who are homeless (Harris & Fiske, 2006, 2009), or people who have settled status versus people seeking asylum (Angelis, 2020; Haslam & Holland, 2012)? The studies in this thesis provide a paradigm to investigate consequences of dehumanisation on language production.
We might also want to continue considering how group identification modulates a Like Me bias. In Experiments 1 and 4, we tested whether a person who identifies more strongly with their gender or racial ingroup would demonstrate a stronger Like Me gender or race bias than a person who identifies less strongly with their ingroup, but we did not find strong evidence to support this. However, that does not necessarily mean ingroup identification plays no role; future investigations could benefit from using other scales than the Multicomponent In-group Identification Survey (Leach et al., 2008), which may be better suited towards testing specific features of identity. It is also possible that our studies were not well-powered enough to detect small effects of individual differences, as we had to exclude many participants for not answering all of the questions on the survey. We might find that robust findings emerge with a larger participant sample.

The ability to measure judgements of agency from mention order could be highly relevant and useful to the study of impression formation, attitudes and person perception. For instance, we might consider applying this paradigm to investigations of victim-blaming. As Bohner (2001) has suggested, first mention is regularly used in descriptions of sexual violence through the passive tense in a way that foregrounds the victim and places the perpetrator in the background. This could, on one hand, provide a necessary focus on the harm suffered by the victim, but on the other it has been found to shift perceptions of blame onto the victim rather than the abuser. Researchers could use our rating paradigm to empirically test just how impactful social biases of mention order are, and from there, work to find solutions to mitigate against biased semantic inferences.
5.2.4. **Societal implications**

Our findings also have important, real-world implications. In establishing that social biases exist in mention order, with consequences for the comprehender, we could make the argument that language, such as that used in reporting or literature, could be used to change the opinions or perceptions of commonly marginalised people, such as people without homes, people seeking asylum, or people who have experienced violence. For instance, when describing violent interactions such as rape, speakers (and writers) should be aware of the assumptions they are creating through who they choose to mention first, so as not to influence negative outcomes such as victim-blaming (Bohner, 2001). In society, words matter for how we structure cultural narratives.

Earlier, I gave the example of The Guardian running stories on “The Russia-Ukraine war” and noted the finding that the first-mentioned group in a conflict is often presumed to be the one with the most power or importance (Oeberst & Matschke, 2017). This is one case where journalists and editors ought to seriously consider the role they are playing in constructing narratives, through who they have chosen to mention first. Perhaps it is time to have a formal policy on first mention across communicative platforms such as those in the media.

It necessary to continue identifying the domains that are likely to reproduce social mention order biases, in order to ascertain where changes should be made. In educational texts where women and men are mentioned together, men are often mentioned first (Ahmad & Shah, 2019). Would discourse analyses would reveal similar patterns in literature, journalism, and other media? And as mentioned in Chapter 2, the majority of journalists in the UK are from white ethnic backgrounds (Spilsbury, 2021), so we might question how this could impact how journalists describe events involving interactions between people of different races. Gender representation is more balanced
in journalism than race (Spilsbury, 2021), but even so, we may consider the possibility
that a woman reporting on an interaction between a woman and a man may do so
differently than a man, and that this in turn could affect the framing of that event.

We can also consider another important domain through which social biases are
regularly detected. As time goes on, increasing numbers of companies are relying on
algorithms to aid in decision-making and to engage with users (Levy et al., 2021).
Algorithmic bias occurs when there are biases in the datasets used to train the
algorithms (Caliskan et al., 2017; Kiritchenko & Mohammad, 2018; Nadeem et al.,
2020). In Natural Language Processing, such biases have been discovered with word
associations, for example associating “homemaker” with the pronoun “she” and
“philosopher” with the pronoun “he” (Bolukbasi et al., 2016) and with sentence
completion, for example finishing the sentence “Two Muslims walked into a” with a
violent sentiment (Abid et al., 2021). From research in this domain, we can expect that
if social biases exist in language, these are likely to be reproduced by machine learning
algorithms. However, psycholinguistics has a long way to go in revealing the diversity
of social language biases. Their identification is therefore relevant for our understanding
of everyday language, but also for predicting and mitigating against biases in algorithms
trained by language models.

5.2.5. **Limitations**

Our language production studies all used data from written language as typed on
a keyboard. Consequently, there is a possibility that our results, most importantly being
whose name was mentioned first, could differ for verbal responses. To test the
ecological validity of text descriptions as reflective of spoken descriptions, we would
have to replicate these findings with spoken language. However, previous research on
conceptual accessibility has shown that similar patterns for spoken language occur in written language (e.g., Rissman et al., 2019; Tachihara & Goldberg, 2020). This suggests that we, too, would find comparable effects of a Like Me bias for verbal responses.

Furthermore, our studies tested a relatively limited range of events. Ten symmetrical verbs were repeated throughout each experiment, many of which were used in all six language-focused experiments. There is scope to test a much wider range of verbs, as well as testing valence in a more nuanced way, for instance through pre-testing and using more formalised categories either within or outside of intimate, casual and negative events. For instance, investigations could address verbs associated with themed events, such as workplace- or employment-related verbs, or with family-oriented events. Thus, testing a wider variety of verbs and events would be beneficial in the context of investigating effects of conceptual accessibility and relative similarity on mention order.

Relying on verbs and actions that imply some symmetry between two people limits the selection slightly, however continuing to test adults would allow investigations to include more complex and uncommon vocabulary. Equally, the verbs used in these studies are, for the most part, appropriate for experiments with child participants. Importantly, this would allow us to replicate our studies with children, to investigate social biases on mention order from a developmental perspective. At what age do children begin to show mention order biases in their production, and biased judgements of agency from semantic inferences? Hsiao and colleagues (2021) have shown that children’s literature is skewed towards representing male characters over female characters, linking this to the overrepresentation of male authors in this area of publishing. Importantly, this bias is reflected in how children themselves write creatively. Hsiao et al. have found that, from the age of five, children are already
demonstrating an own-gender trend in the way they writing, tending to write more about characters who are the same gender as themselves. As they get older, boys continue to show a male preference, whereas girls show a much more equal use of male and female names. But is this a sign of more balance for girls, or this shift towards a male preference in the early stages of what eventually becomes a man-first bias in binomials for both adult men and women as shown by Iliev and Smirnova (2016) Hegarty and colleagues (2011) and Kesebir (2017)? Investigating the development of social biases of mention order could be an important step towards mitigating against these biases, for instance through gender representations in children’s literature, and a more equal society.

5.2.6. Methodological contributions

This thesis offers some important methodological contributions for language research. Firstly, we addressed a common problem in the domain of picture-description tasks, whereby speakers tend to describe images in the order in which they usually read (Maass & Russo, 2003). So, for an English speaker such as those in the experiments of this thesis, this order would be from left to right. This presents a common confound, which is usually only mitigated through counterbalancing which entity appears on the left and right in stimuli. However, the effect of this orientation bias is often so pronounced that it can make it difficult to identify other effects that may be small but meaningful.

This is certainly what we found when we piloted Experiment 1 using still images. We found that participants mentioned the person on the left first around 99% of the time, regardless of the person’s gender or race. Anecdotally, pilot participants also reported that they found themselves simply saying the name of the person on the left
first. After switching to using GIFs, whereby the image switched with its Left-Right mirror image twice per second, we found a great reduction in the orientation bias from 99% to 54%. We found similar results for each subsequent language production experiment. GIFs are easily created using the R package “magick” (Ooms, 2018), requiring no additional stimuli as a mirrored copy of the original image is straightforward to make. We strongly recommend this method for similar picture-description tasks.

Next, we offer our production paradigm, whereby participants in essence “fill in the gaps” using names to complete a sentence, to the study of mention order. Conditions can be easily controlled, as participants are given the exact syntactic structure they should use, and names can either be taught beforehand as in our memory task, or given spontaneously depending on the investigation. This works particularly well for online language studies, both for verbal responses when audio input is possible, and written responses where it may not be possible to collect audio data. Visual cues of the sentence fragment provide a helpful prompt to participants, reducing the likelihood of incorrect or invalid responses and thus excludable data.

Lastly, we have successfully modified the similarity rating scale used by Gleitman et al. (1996) to effectively test judgements of agency. This method could be applied to many different types of sentences and events, as well as for testing properties outside of agency.

5.3. Conclusion

Through these experiments, we have developed and tested an important theory of how social factors can affect mention order, revealing relative social similarity as an
important determinant of first mention. This work reminds us that social cognition is complex, as instead of showing that descriptions of interactions are affected either solely by a Like Me bias or by conventional social hierarchies, our findings suggest that both factors can have an influence on mention order, depending on the social context (demonstrated in finding a man-first bias but not a white-first bias). Furthermore, we have provided quantifiable evidence for the effect of mention order on perceptions of agency and have shown that the first-mentioned agency bias is subject to stereotypes of agency, namely those that are related to gender expectations.

The research in this thesis would not have been possible without taking a multidisciplinary approach to theorising and experimental design. We used principles from psycholinguistics to understand how speakers construct their utterance and how comprehenders infer meaning. Just as importantly, we used principles from social psychology and social cognition to understand how social hierarchies and group identity can affect production and comprehension. Ultimately, this work highlights the speaker as a social agent, whose own biases, assumptions, and perceptions are highly likely to affect how they produce and process language. A speaker’s identity should always be considered in research dedicated to understanding how language is used to talk about and process social events.
6. References


https://doi.org/10.1038/s42256-021-00359-2


https://doi.org/10.1016/j.psychres.2013.09.013


https://doi.org/10.17583/rimcis.2019.3989


https://doi.org/10.1023/A:1024109008573


https://doi.org/10.1017/S1474746419000216


https://doi.org/10.1016/j.cognition.2009.10.015


Buckle, L. (2020). *A Developmental Investigation of Animacy and Prototypical Transitivity in Sentence Production and Comprehension* [Ph.D., The University of Manchester (United Kingdom)].

https://www.proquest.com/docview/2564454480/abstract/511DCE1BB8C6481CPQ/1


Douglass, P. D. (2020). *Assata is Here: (Dis)Locating Gender in Black Studies*. https://www-tandfonline-


   http://dx.doi.org/10.1057/s41305-017-0088-1


   https://doi.org/10.2307/413059


   https://doi.org/10.1016/j.bandl.2016.03.002


Oeberst, A., & Matschke, C. (2017). Word order and world order: Titles of intergroup conflicts may increase ethnocentrism by mentioning the in-group first. *Journal*
of Experimental Psychology: General, 146(5), 672–690.
https://doi.org/10.1037/xge0000300

https://doi.org/10.1016/j.cogpsych.2007.04.001


https://doi.org/10.1111/desc.12788

https://doi.org/10.1080/00207599308246932

https://doi.org/10.1006/jmla.1999.2668


https://doi.org/10.1177/0963721417694353


https://doi.org/10.1371/journal.pone.0146359
7. Appendix A: Experimental stimuli used in Experiments 1-7

7.1. Experimental stimuli used in Experiment 1

Table A1: List of phrases used in Experiment 1.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Transitive</th>
<th>Conjoined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argue</td>
<td>is arguing with</td>
<td>are arguing with each other</td>
</tr>
<tr>
<td>Dance</td>
<td>is dancing with</td>
<td>are dancing with each other</td>
</tr>
<tr>
<td>Fight</td>
<td>is fighting with</td>
<td>are fighting each other</td>
</tr>
<tr>
<td>Hug</td>
<td>is hugging</td>
<td>are hugging each other</td>
</tr>
<tr>
<td>Kiss</td>
<td>is kissing</td>
<td>are kissing each other</td>
</tr>
<tr>
<td>Marry</td>
<td>is marrying</td>
<td>are marrying each other</td>
</tr>
<tr>
<td>Meet</td>
<td>is meeting</td>
<td>are meeting each other</td>
</tr>
<tr>
<td>Shout</td>
<td>is shouting at</td>
<td>are shouting at each other</td>
</tr>
<tr>
<td>Talk</td>
<td>is talking to</td>
<td>are talking to each other</td>
</tr>
<tr>
<td>Touch</td>
<td>is touching</td>
<td>are touching each other</td>
</tr>
</tbody>
</table>
Figure A1: Matrix of actions used in Experiment 1.
### 7.2. Experimental stimuli used in Experiment 2, 3 and 5

#### Table A2: List of phrases used in Experiments 2 and 5.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Transitive</th>
<th>Conjoined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argue</td>
<td>is arguing with</td>
<td>are arguing with each other</td>
</tr>
<tr>
<td>Dance</td>
<td>is dancing with</td>
<td>are dancing with each other</td>
</tr>
<tr>
<td>Fight</td>
<td>is fighting with</td>
<td>are fighting each other</td>
</tr>
<tr>
<td>Flirt</td>
<td>is flirting with</td>
<td>are flirting with each other</td>
</tr>
<tr>
<td>Kiss</td>
<td>is kissing</td>
<td>are kissing each other</td>
</tr>
<tr>
<td>Marry</td>
<td>is marrying</td>
<td>are marrying each other</td>
</tr>
<tr>
<td>Play</td>
<td>is playing with</td>
<td>are playing with each other</td>
</tr>
<tr>
<td>Shout</td>
<td>is shouting at</td>
<td>are shouting at each other</td>
</tr>
<tr>
<td>Talk</td>
<td>is talking to</td>
<td>are talking to each other</td>
</tr>
<tr>
<td>Touch</td>
<td>is touching</td>
<td>are touching each other</td>
</tr>
</tbody>
</table>

#### Table A3: List of rating questions used in Experiments 3 and 5. The inserted image (Experiment 3) or name (Experiment 5) is represented by an underscore.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Causation</th>
<th>Volition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argue</td>
<td>_ started an argument with _</td>
<td>_ wanted to argue with _</td>
</tr>
<tr>
<td>Dance</td>
<td>_ asked _ to dance</td>
<td>_ wanted to dance with _</td>
</tr>
<tr>
<td>Fight</td>
<td>_ started the fight with _</td>
<td>_ wanted to fight _</td>
</tr>
<tr>
<td>Flirt</td>
<td>_ started flirting with _ first</td>
<td>_ wanted to flirt with W_</td>
</tr>
<tr>
<td>Action</td>
<td>asked to</td>
<td>wanted to</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>Kiss</td>
<td>_ kiss</td>
<td>_ kiss</td>
</tr>
<tr>
<td>Marry</td>
<td>_ marry</td>
<td>_ marry</td>
</tr>
<tr>
<td>Play</td>
<td>_ play</td>
<td>_ play</td>
</tr>
<tr>
<td>Shout</td>
<td>_ shout</td>
<td>_ shout</td>
</tr>
<tr>
<td>Talk</td>
<td>_ talk</td>
<td>_ talk</td>
</tr>
<tr>
<td>Touch</td>
<td>_ touch</td>
<td>_ touch</td>
</tr>
</tbody>
</table>
Figure A2: Matrix of actions used in Experiments 2 and 3.

<table>
<thead>
<tr>
<th>Argue</th>
<th>Dance</th>
<th>Fight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flirt</td>
<td>Kiss</td>
<td>Marry</td>
</tr>
<tr>
<td>Play</td>
<td>Shout</td>
<td>Talk</td>
</tr>
<tr>
<td>Touch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 7.3. Experimental stimuli used in Experiment 4

Table A4: List of phrases used in Experiment 4.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Active</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenge</td>
<td>is challenging</td>
<td>is being challenged by</td>
</tr>
<tr>
<td>Copy</td>
<td>is copying</td>
<td>is being copied by</td>
</tr>
<tr>
<td>Greet</td>
<td>is greeting</td>
<td>is being greeted by</td>
</tr>
<tr>
<td>Hug</td>
<td>is hugging</td>
<td>is being hugged by</td>
</tr>
<tr>
<td>Kiss</td>
<td>is kissing</td>
<td>is being kissed by</td>
</tr>
<tr>
<td>Look</td>
<td>is looking at</td>
<td>is being looked at by</td>
</tr>
<tr>
<td>Oppose</td>
<td>is opposing</td>
<td>is being opposed by</td>
</tr>
<tr>
<td>Seduce</td>
<td>is seducing</td>
<td>is being seduced by</td>
</tr>
<tr>
<td>Shout</td>
<td>is shouting at</td>
<td>is being shouted at by</td>
</tr>
<tr>
<td>Touch</td>
<td>is touching</td>
<td>is being touched by</td>
</tr>
</tbody>
</table>
Figure A3: Matrix of actions used in Experiment 4.
### 7.4. Experimental stimuli used in Experiments 6 and 7

Table A6: List of phrases used in Experiments 6 and 7.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Sentence fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argue</td>
<td>is arguing with</td>
</tr>
<tr>
<td>Dance</td>
<td>is dancing with</td>
</tr>
<tr>
<td>Fight</td>
<td>is fighting with</td>
</tr>
<tr>
<td>Flirt</td>
<td>is flirting with</td>
</tr>
<tr>
<td>Kiss</td>
<td>is kissing</td>
</tr>
<tr>
<td>Marry</td>
<td>is marrying</td>
</tr>
<tr>
<td>Play</td>
<td>is playing with</td>
</tr>
<tr>
<td>Shout</td>
<td>is shouting at</td>
</tr>
<tr>
<td>Talk</td>
<td>is talking to</td>
</tr>
<tr>
<td>Touch</td>
<td>is touching</td>
</tr>
</tbody>
</table>
Figure A5: Matrix of actions used in Experiment 6 and 7.

Argue  Dance  Fight

Flirt  Kiss  Marry

Play  Shout  Talk

Touch
Table A7: Short Stories and instructions used in Experiment 7, including the first image participants viewed as a device to set the scene. Differences in wording between the narrative conditions are represented in *italics*, experience cues are represented with *underlining*, and agency cues are represented with **bold**. These formatting styles were not present in the display of experiment stimuli.

<table>
<thead>
<tr>
<th>Human-oriented</th>
<th>Robot-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
</tbody>
</table>

In a distant future, many nations on Planet Teran, populated by *fleshy*, sentient beings, have developed their own Space Military Department. The Global Space Force Movement, which unites them all, now boasts the most rigorous in-space training programme in history, with meticulously trained armed forces utilising weapons that work in all atmospheres and orbits.

The Space Military now seeks to expand its defensive armed forces by developing a Secret Space Intelligence and Defence Service, creating a team for counterintelligence, espionage and security that can infiltrate foreign planets posing a risk to Teran.

Their current greatest perceived threat is from Planet Trill – a planet about 7 billion km...
away from Teran, populated by an intelligent, \textit{bio-metallic} species. Scientists on Teran have detected worrying signals coming from Trill, including atomic radiation waves that are usually only emitted from weapons of mass atmospheric destruction.

Four Biologists have been recruited to join the Secret Space Intelligence and Defence Service, as part of the espionage force. Their task is to grow clones that are perfectly modelled on the population of Trill. These clones, or ‘Spy-Bods’ will infiltrate Trill’s military to spy on those working on the threatening weaponry detected by Teran’s Scientists. They will then transmit the gathered information back to Teran’s Headquarters.

\textbf{Meet the Biologists}

Biologist Cybe has been working in the field of cloning for several years and is \textit{happiest} when in the lab. Cybe has a PhD in psychosocial interactions and \textit{enjoys} visiting museums on the weekend.
<table>
<thead>
<tr>
<th>Biologist Mech is passionate about stem cell recreation and is a proud senior researcher with the Scientific Social Research Council. Mech speaks several languages and has two pets.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biologist Botz teaches as well as doing research and is a professor at a prestigious educational institution. Botz’s life mission is to reduce pollution from scientific research and can often be found enjoying an environmental activism march.</td>
</tr>
<tr>
<td>Biologist Roid is an award-winning researcher whose speciality lies in developmental physiology. Roid also loves to compose music and knows how important it is for scientists to have multiple creative</td>
</tr>
<tr>
<td>is for scientists to have multiple creative outlets.</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>The Biologists must produce flawless spying clones. The clones must have expertly-trained Response And React functions that mimic the language and communication diversity of Trill’s beings. These Spy-Bods must also have the ability to carry out fine-tuned movements and facial expressions. They cannot be detectable as foreign, otherwise all the progress of the Defence operation thus far will be destroyed.</td>
</tr>
<tr>
<td><strong>After working together for months, the Biologists have reached the first testing stage of this project: they have grown some prototype Spy-Bods, which they have trained to mimic actions with a partner. The Biologists are required to show that their Spy-Bods, which they have given common Trill names, can perform a selection of tasks and social interactions – a test you are about to witness.</strong></td>
</tr>
<tr>
<td>The Biologists are nervous about the assessment, but they think their Spy-Bods are ready. They are planning to impress the Space Defence team by interacting with the</td>
</tr>
</tbody>
</table>
Spy-Bods themselves and demonstrating just how natural these interactions look.
Most of all, the Biologists want to help protect their beloved planet from attacks.

The Spy-Bods

Your role is to describe the interactions taking place during this test. We will give you detailed instructions shortly. Press ‘Continue’.

On the next page, you will see portraits of the 4 Biologists and the 4 Spy-Bods, and their names again.
To carry out this task, you must first take a few minutes to memorise the names of all the Biologists and all the Spy-Bods.

You will have to recall these names during the rest of the task, so it is essential that you try to memorise these portraits and names as best as you can.

You will then be tested on your memory of this information.

The task will start after you have completed this memory test.

Ready?
If you have memorised all the portraits and names and are ready to start, press Continue. Please note: Your memory of these portraits will be tested 4 times during this experiment.

You are about to view photographs of the Biologists and Spy-Bods interacting, as part of the Biologists’ test.

By using the exact phrase given after the photograph appears and recalling the correct names of the figures in the photographs, form a sentence describing the interaction. You do not need to capitalise names.

Type your answer in the space provided. Please complete every sentence before moving onto the next trial.

You will only have a few seconds to memorise each given photograph and phrase before you can respond, so try your best to focus.

An example is given below. Press 'Continue' when you are ready.

Before the test starts, you will watch two example trials. Press 'Continue'.

Ready to start the Biologists’ test?

Remember to describe the picture using a full sentence, that combines the names of the two figures with the exact phrase given.
To receive full payment, you must do this correctly.

You do not need to capitalise names or use punctuation.

Press 'Continue' to begin.

Experience cues in text

“happiest” = joy
“enjoys” = pleasure
“passionate” = joy
“proud” = pride
“life mission” = desire
“enjoying” = pleasure
“loves” = pleasure
“knows” = consciousness
“nervous” = fear
“want” = desire

All “Biologist” profiles cue experience property ‘personality’

Agency cues in text

“After working together for months” = communication
“they think” = thought
“They are planning to impress” = planning
“want to help protect their beloved planet from attacks” = morality
8. Appendix B: Scales used for questionnaires in Experiments 1 – 7

8.1. Items Measuring In-Group Identification for gender and race

Based on ‘Group-Level Self-Definition and Self-Investment: A Hierarchical (Multicomponent) Model of In-Group Identification (Leach et al., 2008).

(Group-Level) Self-Investment

Solidarity

0. I feel a bond with [my gender/my race].
1. I feel solidarity with [my gender/my race].
2. I feel committed to [my gender/my race].

Satisfaction

3. I am glad to be [my gender/my race].
4. I think that [my gender/my race] has a lot to be proud of.
5. It is pleasant to be [my gender/my race].

Centrality

7. I often think about the fact that I am [my gender/my race].
8. The fact that I am [my gender/my race] is an important part of my identity.
9. Being [my gender/my race] is an important part of how I see myself.

(Group-Level) Self-Definition

Individual Self-Stereotyping

10. I have a lot in common with the average person of [my gender/my race].
11. I am similar to the average person of [my gender/my race].

In-group Homogeneity

12. People of [my gender/my race] have a lot in common with each other.
13. People of [my gender/my race] are very similar to each other.
8.2. Items measuring attitudes towards same-gender relationships and interracial relationships

Items testing attitudes towards same-gender relationships with references.

1. Gay men and lesbians should be free to live their own life as they wish.

"Strongly Disagree", "Disagree", "Neither Agree nor Disagree", "Agree", "Strongly Agree"

(ESS 2002–2010)

2. What are your opinions about sexual relations between two adult men?
3. What are your opinions about sexual relations between two adult women?

“Always Wrong”, “Mostly Wrong”, “Sometimes Wrong”, “Rarely Wrong”, “Not Wrong At All”

(Watt & Elliot, 2019)

Attitudes towards interracial relationships

4. I think it is good for Black and White people to be friends.
5. I think it is good for Black and White people to date.
6. I think it is good for Black and White people to marry.

"Strongly Disagree", "Disagree", "Neither Agree nor Disagree", "Agree", "Strongly Agree"

(Field et al., 2013)
8.2. **Items Measuring Attitudes towards Robots**

Negative attitudes towards Robots Scale (NARS, Nomura et al., 2008). Response format: five-point scale:

“Strongly disagree”, “Disagree”, “Undecided”, “Agree”, “Strongly agree”

Subscale 1: Negative attitude toward interaction with robots (6 items, $\alpha = .738$)

1. I would feel uneasy if I was given a job where I had to use robots.
2. The word “robot” means nothing to me.
3. I would feel nervous operating a robot in front of other people.
4. I would hate the idea that robots or artificial intelligences were making judgments about things.
5. I would feel very nervous just standing in front of a robot.
6. I would feel paranoid talking with a robot.

Subscale 2: Negative attitude toward the social influence of robots (5 items, $\alpha = .732$)

7. I would feel uneasy if robots really had emotions.
8. Something bad might happen if robots developed into living beings.
9. I feel that if I depend on robots too much, something bad might happen.
10. I am concerned that robots would be a bad influence on children.
11. I feel that in the future society will be dominated by robots.

Subscale 3: Negative attitude toward emotional interactions with robots (3 items, $\alpha = .657$)

12. I would feel relaxed talking with robots. (reverse)
13. If robots had emotions, I would be able to make friends with them. (reverse)
14. I feel comforted being with robots that have emotions. (reverse)
8.3. Items for Manipulation Check (Dimensions of Mind Perception and Dehumanisation scale)

Two separate questionnaires were used for the human-oriented and robot-oriented conditions, as the group the dehumanisation questions referred to depended on condition (to the human “Biologist” group in the human-oriented condition, and to the human “Spy-Bod” group in the robot-oriented condition). For items from the Dimensions of Mind Perception, items that were specifically cued in the narrative are marked in bold. For each sub-scale, the response scales ranged from 1 = not at all to 100 = extremely.

8.3.1. Robot-Oriented Condition

“Biologists” = robot group
“Spy-Bods” = human group

Dimensions of Mind Perception: items taken from Gray et al. (2007).

Experience

1. Hunger
   - To what extent are [Biologists/Spy-Bods] capable of feeling hungry.
2. Fear
   - To what extent are [Biologists/Spy-Bods] capable of feeling afraid or fearful.
3. Pain
   - To what extent are [Biologists/Spy-Bods] capable of experiencing physical or emotional pain.

4. Pleasure
   - To what extent are [Biologists/Spy-Bods] capable of experiencing physical or emotional pleasure.

5. Rage
   - To what extent are [Biologists/Spy-Bods] capable of experiencing violent or uncontrolled anger.
6. Desire
   - To what extent are [Biologists/Spy-Bods] capable of longing or hoping for things.

7. Personality
   - To what extent are [Biologists/Spy-Bods] capable of having personality traits that make them unique from others.

8. Consciousness
   - To what extent are [Biologists/Spy-Bods] capable of having experiences and being aware of things.

9. Pride
   - To what extent are [Biologists/Spy-Bods] capable of experiencing pride.

10. Embarrassment
   - To what extent are [Biologists/Spy-Bods] capable of experiencing embarrassment.

11. Joy
   - To what extent are [Biologists/Spy-Bods] capable of experiencing joy.

Agency

12. Self-control
   - To what extent are [Biologists/Spy-Bods] capable of exercising self-restraint over desires, emotions, or impulses.

13. Morality
   - To what extent are [Biologists/Spy-Bods] capable of telling right from wrong and trying to do the right thing.

14. Memory
   - To what extent are [Biologists/Spy-Bods] capable of remembering things.

15. Emotion Recognition
   - To what extent are [Biologists/Spy-Bods] capable of understanding how others are feeling.

16. Planning
   - To what extent are [Biologists/Spy-Bods] capable of making plans and working toward goal.

17. Communication
- To what extent are [Biologists/Spy-Bods] capable of conveying thoughts or feelings to others.

18. Thought

- To what extent are [Biologists/Spy-Bods] capable of thinking.

Empathy: items 19 – 21 are from Seo et al. (2015), 22 and 23 are original items.

19. How worried are you for the [Biologists/Spy-Bods]?
20. How moved are you by the [Biologists/Spy-Bods]?
21. How sympathetic do you feel towards the [Biologists/Spy-Bods]?
22. How empathetic do you feel towards the [Biologists/Spy-Bods]?
23. How animate are the [Biologists/Spy-Bods]?

Dehumanisation: items taken from Harris and Fiske (2011).

General traits

24. How competent are the Spy-Bods?
25. How warm are the Spy-Bods?
26. How similar to you are the Spy-Bods?
27. How familiar to you are the Spy-Bods?
28. How articulate are the Spy-Bods?
29. How likely are the Spy-Bods to experience the ups and downs of life?
30. How typically human are the Spy-Bods?
31. How intelligent are the Spy-Bods?

Story traits

I felt like the Spy-Bods in the description were….

32. Open minded, like they could think clearly about things.
33. Emotional, like they were responsive and warm.
34. Superficial like they had no depth. (reverse-coded)
35. Mechanical and cold, like a robot. (reverse-coded)
36. Refined and cultured.
37. Rational and logical, like they were intelligent.
38. Lacked self-restraint, like an animal. (reverse-coded)
39. Unsophisticated. (reverse-coded)

8.3.2. Human-Oriented Condition

“Biologists” = human figures
“Spy-Bods” = robot figures

Dimensions of Mind Perception: items taken from Gray et al. (2007).

**Experience**

1. Hunger
   - To what extent are Biologists/Spy-Bods capable of feeling hungry.

2. **Fear**
   - To what extent are Biologists/Spy-Bods capable of feeling afraid or fearful.

3. Pain
   - To what extent are Biologists/Spy-Bods capable of experiencing physical or emotional pain.

4. **Pleasure**
   - To what extent are Biologists/Spy-Bods capable of experiencing physical or emotional pleasure.

5. Rage
   - To what extent are Biologists/Spy-Bods capable of experiencing violent or uncontrolled anger.

6. **Desire**
   - To what extent are Biologists/Spy-Bods capable of longing or hoping for things.

7. **Personality**
   - To what extent are Biologists/Spy-Bods capable of having personality traits that make them unique from others.
8. Consciousness
   - To what extent are Biologists/Spy-Bods capable of having experiences and being aware of things.

9. Pride
   - To what extent are Biologists/Spy-Bods capable of experiencing pride.

10. Embarrassment
    - To what extent are Biologists/Spy-Bods capable of experiencing embarrassment.

11. Joy
    - To what extent are Biologists/Spy-Bods capable of experiencing joy.

Agency

12. Self-control
    - To what extent are Biologists/Spy-Bods capable of exercising self-restraint over desires, emotions, or impulses.

13. Morality
    - To what extent are Biologists/Spy-Bods capable of telling right from wrong and trying to do the right thing.

14. Memory
    - To what extent are Biologists/Spy-Bods capable of remembering things.

15. Emotion Recognition
    - To what extent are Biologists/Spy-Bods capable of understanding how others are feeling.

16. Planning
    - To what extent are Biologists/Spy-Bods capable of making plans and working toward goal.

17. Communication
    - To what extent are Biologists/Spy-Bods capable of conveying thoughts or feelings to others.

18. Thought
    - To what extent are Biologists/Spy-Bods capable of thinking.
Empathy: items 19 – 21 are from Seo et al. (2015), 22 and 23 are original items.

19. How worried are you for the Biologists?
20. How moved are you by the Biologists?
21. How sympathetic do you feel towards the Biologists?
22. How empathetic do you feel towards the Biologists?
23. How animate are the Biologists?

Dehumanisation: items taken from Harris and Fiske (2011).

General traits

24. How competent are the Biologists?
25. How warm are the Biologists?
26. How similar to you are the Biologists?
27. How familiar to you are the Biologists?
28. How articulate are the Biologists?
29. How likely are the Biologists to experience the ups and downs of life?
30. How typically human are the Biologists?
31. How intelligent are the Biologists?

Story traits

I felt like the Biologists in the description were….

32. Open minded, like they could think clearly about things.
33. Emotional, like they were responsive and warm.
34. Superficial like they had no depth (reverse-coded)
35. Mechanical and cold, like a robot. (reverse-coded)
36. Refined and cultured.
37. Rational and logical, like they were intelligent.
38. Lacked self-restraint, like an animal (reverse-coded)
39. Unsophisticated (reverse-coded)