Synthese (2021) 199:1347–1385
https://doi.org/10.1007/s11229-020-02798-7

THE CULTURAL EVOLUTION OF HUMAN SOCIAL COGNITION

A computational model of the cultural co-evolution of language and mindreading

Marieke Woensdregt[1] · **Chris Cummins**[2] · **Kenny Smith**[3]

Received: 9 August 2019 / Accepted: 14 July 2020 / Published online: 2 October 2020

Abstract

Several evolutionary accounts of human social cognition posit that language has co-evolved with the sophisticated mindreading abilities of modern humans. It has also been argued that these mindreading abilities are the product of cultural, rather than biological, evolution. Taken together, these claims suggest that the evolution of language has played an important role in the cultural evolution of human social cognition. Here we present a new computational model which formalises the assumptions that underlie this hypothesis, in order to explore how language and mindreading interact through cultural evolution. This model treats communicative behaviour as an interplay between the context in which communication occurs, an agent's individual perspective on the world, and the agent's lexicon. However, each agent's perspective and lexicon are private mental representations, not directly observable to other agents. Learners are therefore confronted with the task of jointly inferring the lexicon and perspective of their cultural parent, based on their utterances in context. Simulation results show that given these assumptions, an informative lexicon evolves not just under a pressure to be successful at communicating, but also under a pressure for accurate perspective-inference. When such a lexicon evolves, agents become better at inferring others' perspectives; not because their innate ability to learn about perspectives changes, but because sharing a language (of the right type) with others helps them to do so.

Keywords Language evolution · Mindreading · Cultural evolution · Computational modelling · Iterated learning · Bayesian inference

M.W. was funded by a Principal's Career Development Ph.D. Scholarship from the University of Edinburgh.

✉ Marieke Woensdregt
m.woensdregt@let.ru.nl

[1] Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

[2] Department of Linguistics and English Language, University of Edinburgh, Edinburgh, UK

[3] Centre for Language Evolution, University of Edinburgh, Edinburgh, UK

1 Introduction

The hypothesis that human social cognition is a product of cultural evolution is motivated in part by evidence showing that language plays a role in the development of social cognition in modern-day humans (Heyes 2018; Heyes and Frith 2014). Studies of typically-developing children (Astington and Baird 2005; Milligan et al. 2007), and deaf children who had a delayed exposure to sign language (see Meristo et al. 2011; Peterson and Siegal 2000; Pyers and de Villiers 2013, for reviews) have shown that exposure to mental state language and discourse promotes the development of mental state attribution (i.e. *mindreading*; also known as *theory of mind*). Furthermore, longitudinal studies have shown that caregivers tailor *how* they talk about mental states to the developmental stages of their children, in a way that helps children learn about other minds (see Slaughter and Peterson 2011; Meins 2011, for reviews). Taken together, these findings indicate that cultural learning—in the sense of social learning that makes use of cognitive processes specialised for cultural inheritance, such as language (Heyes 2018, pp. 86-87)—plays a role in the development of mindreading. If it is the case that language supports and promotes our thinking about other minds, as well as the cultural transmission of such thinking, then the emergence and evolution of language in hominins may have unlocked levels of mindreading ability that could not have been attained without it.

The other way around, it has also been argued that the evolution of more sophisticated mindreading abilities in hominins has played an important role in the evolution of language. Language is a form of *Gricean* (also known as *ostensive-inferential*) communication, and the level of mindreading sophistication that is minimally required to embark on such communication is a matter of debate (Moore 2017b; Scott-Phillips 2015a), as is the question whether our closest living relatives, the nonhuman great apes, possess these minimal requirements (Moore 2016, 2017a, c; Scott-Phillips 2015a, b, c, 2016; Townsend et al. 2017). However, there is no doubt that the explicit mindreading skills we find in humans today, which seem to exceed those of nonhuman great apes (Call and Tomasello 2008; Heyes 2015, 2017; Martin and Santos 2016), support our language use. Language use involves a speaker tailoring their utterances to (their model of) the hearer's mind, and the hearer in turn making inferences about the speaker's mind when interpreting those utterances (Grice 1957; Moore 2017b; Scott-Phillips 2015a; Sperber and Wilson 1995).

In this paper we demonstrate how computational modelling can be used to explore the hypothesis that the emergence and cultural evolution of language has led to better mindreading. Computational modelling provides a way of formalising the assumptions about learning, development and cultural transmission that underlie this hypothesis. This allows us to explore a potential evolutionary scenario that represents a middle ground between the two perspectives on the evolutionary interdependence between mindreading and language described above: the hypothesis that language and mindreading have co-evolved (see e.g. Malle 2002; Whiten and Erdal 2012). Under what circumstances could such a co-evolution have gotten off the ground? And could this happen by means of cultural evolution?

In Sect. 2 we briefly review empirical evidence of the role that language plays in the development of mindreading, and vice versa, and discuss how this informs our

model of learning. In Sect. 3, we review different theoretical views on the evolution of language and mindreading and the extent to which one builds on the other. In Sect. 4 we briefly review existing computational models of word learning and language evolution that form the basis for the agent-based model that we present in this paper. In Sect. 5 we go on to describe this model, in which individuals infer both a lexicon and their interlocutor's perspective, based on observable behaviour (word use in context), where neither the lexicon nor the perspective can be observed directly. This model allows us to explore how language and perspective-taking interact on a cultural evolutionary timescale. We present the evolutionary dynamics that ensue in Sect. 6, where we show that the cultural evolution of an informative lexicon can improve agents' success at inferring others' perspectives, without agents' underlying 'innate' ability to learn about others' perspectives having changed. The implications of these results for theoretical work on the cultural evolution of human social cognition, and the potential for future modelling work on this topic, are discussed in Sect. 7.

2 Co-development of language and mindreading

Mindreading (also known as *theory of mind*, *mentalising*, or *folk psychology*) refers to the ability to attribute mental states to oneself and others, which is used to explain and predict behaviour. The development of this ability consists of progressive stages of realisation that others' perspectives can differ from one's own: from understanding diverse desires to understanding false beliefs (Apperly 2011; Wellman and Liu 2004; Wellman et al. 2011). False belief understanding is widely taken as a litmus test for a fully-fledged mindreading ability, because it demonstrates an understanding of the 'representational nature' of mind: the ability to represent another individual's belief about the world completely independently from one's own. The staged development of mindreading (and the fact that its order varies somewhat depending on the cultural context[1]; see e.g. Slaughter and Perez-Zapata 2014) suggests that mindreading is a complex skill that may well rely on multiple subsystems. This view has been put forward by 'two-systems' accounts of mindreading (Apperly and Butterfill 2009; Heyes 2018, chapter 7), which distinguish between *implicit* mindreading abilities that are cognitively efficient but inflexible, and *explicit* mindreading abilities that are flexible but cognitively demanding and later to develop. There is empirical evidence that language input plays a role particularly in the development of these explicit mindreading abilities. Conversely, there is also evidence that mindreading plays a role in language development. Below we briefly discuss each of these directions of influence, and how they informed the model that we will go on to present in Sect. 5.

2.1 The role of language in mindreading development

As mentioned above, it has been shown that linguistic input is important for mindreading development. Firstly, longitudinal studies have shown that the extent to which

[1] This in itself is also evidence in favour of the hypothesis that mindreading is (at least in part) culturally transmitted, as discussed above.

caregivers refer to mental states when talking to their child predicts the child's later performance on mindreading tasks (see Meins 2011; Slaughter and Peterson 2011, for reviews). Secondly, the mindreading development of deaf children of hearing parents, who were late to start learning sign language, has been shown to be delayed compared to that of both their hearing and native signing peers (see Meristo et al. 2011; Peterson and Siegal 2000; Pyers and de Villiers 2013, for reviews). This suggests that early conversational interactions are important for mindreading development. Just how important is illustrated by studies of deaf people who grew up without exposure to an existing sign language. Pyers and Senghas (2009) showed that the first cohort of signers of the recently emerged Nicaraguan Sign Language (NSL)—that is, the people who first started to develop the language together when they were children—performed worse on a minimally linguistic false belief task than the second cohort of signers. This was found despite the fact that the first cohort is (on average) ten years older, and had thus had ten more years of general social experience. Two years later however, once signers from the first cohort started using mental state verbs which had been added to the language by the second cohort, their false belief understanding improved. Gagne and Coppola (2017) showed that deaf individuals in Nicaragua who grew up in isolation from a deaf community and thus developed an idiosyncratic gestural communication system to communicate with family members, known as *homesign*, were outperformed on a false belief task by both hearing people and the first cohort of NSL signers. Taken together, these studies indicate that engaging in and observing *non-linguistic* social interactions is not sufficient for developing explicit false belief understanding.

Several hypotheses have been put forward as to what it is about language that makes it helpful for mindreading development (Lohmann and Tomasello 2003; Apperly 2011, pp. 26-30). These are that language provides (i) labels for mental states (Olson 1988), (ii) sentential complement syntax which forms a representational framework that helps to think about beliefs (de Villiers and Pyers 2002), and (iii) conversational experience that highlights differences in perspectives (Harris 2005). These different hypotheses are of course not mutually exclusive, and they are hard to test in isolation because the different aspects of language that they single out tend to go hand-in-hand. Therefore, Lohmann and Tomasello (2003) designed a controlled false belief training study in order to pit the second and third of these hypotheses against each other. They showed that both a 'sentential complement only' training and a 'discourse only' training condition (using deceptive objects) could improve three-year-olds' performance on false belief tasks. Furthermore, they showed that the strongest improvement was achieved in a condition which combined both types of training (while keeping the overall training time constant), indicating that the benefits of sentential complements and perspective-shifting discourse are additive and therefore independent of one another.

The model of cultural evolution we will present below takes as its basis a very simple and abstract model of how language and mindreading co-develop. We model mindreading as the ability to infer another agent's perspective on the world, which, in combination with observing a given situation, allows the learner to predict how likely that agent is to talk about different referents in that situation. In this model, mindreading is the ability to infer another agent's perspective based on observations of their linguistic behaviour in different contexts (i.e. situations). We thus assume that

linguistic input plays a crucial role in the development of mindreading. However, a lot of variation is possible in the quality of this input. Learners can receive input from a very informative language where utterances map unambiguously to individual referents; they can receive input from an entirely ambiguous language where every utterance is associated with every referent; or anything in between. The more informative the language is about its speaker's referential intentions, the easier it will be for the learner to infer that speaker's perspective on the world, because observing what a speaker tends to talk about in different contexts is the learner's only 'way in' to inferring their perspective. This seems like an extreme set-up: in real life children have more data to go on for their mindreading development than just linguistic input. However, it is the role that language plays in reaching full-blown explicit mindreading (as reviewed above) that we are interested in here.

In terms of exactly *how* language aids the development of mindreading, we assume the simplest possible model: receiving input from a useful language does not cause any qualitative change to the way the learner represents the relationship between the context, the speaker's perspective and their utterances. However, the more informative the language is, the easier it will be for the learner to accurately infer the speaker's perspective. Our model thus most closely resembles hypothesis (iii) above: that what is special about language is that it provides conversational experience that highlights differences in perspectives (Harris 2005). Extensions of this model that more closely resemble hypotheses (i) or (ii) above are also possible; we will return to this in the Discussion. In terms of cultural evolution, we are interested in seeing under what circumstances a fully informative language (the type that is helpful for perspective-inference) can emerge from scratch.

2.2 The role of mindreading in language development

In addition to the development of explicit mindreading depending on linguistic input, there is also empirical evidence of mindreading being important for language development. We are interested in this co-development of mindreading and language, and how it plays out on a cultural evolutionary timescale. In the model presented below, we therefore assume not just that perspective-inference depends on the learning of an informative language, but also the other way around: that language-learning depends on accurate perspective-inference. Below we will briefly summarise the theoretical and empirical literature that motivates this assumption.

In general, language use requires a certain degree of mindreading in the sense that a speaker has to do at least some modelling of their audience in order to choose their utterance (e.g. to decide whether to talk about "the house" or "a house"), and a listener has to do some modelling of the speaker in order to recover the speaker's intended meaning, over and above the literal meaning of the utterance (e.g. to infer that "I'm tired" means "Let's go home") (Clark and Carlson 1981; Clark and Marshall 1981; Clark and Murphy 1982; Clark and Wilkes-Gibbs 1986; Grice 1975). Here we will focus on the role of mindreading in language *development* however.

At least from the age of 19 months onwards, infants use social cues such as eye-gaze in order to infer the referent of a novel word (see Baldwin and Moses 2001; Tomasello

2000, for reviews). Using eye-tracking, Yurovsky and Frank (2017) showed that from 1 to 3.5 years old, children become progressively better at following a speaker's eye-gaze to her intended referent, and that this goes together with an improved ability to pick out the target object at test (where test trials used an implicit measure: the preferential looking paradigm). Nappa et al. (2009) showed that 3- to 5-year-olds are able to use a speaker's eye-gaze to learn the meaning of a novel verb in an ambiguous context. This study showed that children can use a speaker's gaze not just to make inferences about an intended referent, but also to make inferences about what a sentence is 'about' (i.e. how the speaker is framing the event). Tomasello and Barton (1994), Akhtar and Tomasello (1996), and Tomasello et al. (1996) showed that 1.5- to 2-year-old children are able to infer a speaker's referential intention even if the novel word never co-occurs with the intended referent, meaning they could not use a direct cue like eye-gaze, but instead had to rely on their inferences about the speaker's intentions. Finally, in an adaptation of the experiment of Tomasello and Barton (1994), Parish-Morris et al. (2007) found that children with autism spectrum disorder (which is characterised in part by an impaired ability to understand social cues and infer others' mental states; American Psychiatric Association 2013; Baron-Cohen et al. 1985; Baron-Cohen 1995) have difficulty learning words when it requires inferring the speaker's intention (because other social cues to reference weren't made available in that experimental condition). Children with autism spectrum disorder are very often delayed in their language development (Eigsti et al. 2011; Tager-Flusberg et al. 2005).

In sum, there is good evidence that at least implicit mindreading (in the form of inferring what another person is attending to, and what their goal is at a given moment) is involved in word learning from a relatively young age (see also Sabbagh and Baldwin 2005), and that if such implicit mindreading is impaired, this causes delays in language development. Once explicit mindreading abilities come online (possibly as a result of exposure to language, as argued in Sect. 2.1 above), it is likely that these are deployed in the same way, especially when it comes to inferring the meaning of words that describe more abstract concepts (see e.g. Papafragou 2002; Ünal and Papafragou 2016).

In the model we present in this paper, we assume that language-learning relies on perspective-inference. We put our learners in a difficult position, where they cannot use cross-situational learning (Siskind 1996) in order to infer the referent of a novel word, nor are they helped by more direct social cues like eye-gaze or pointing. Instead, their only 'way in' to inferring the speaker's most likely referential intention is by inferring their perspective on the world (in combination with observing the current context). Thus, as will be described in more detail in Sect. 5, we assume that learners are confronted with a joint inference task: they have to simultaneously infer both the perspective and the lexicon of the cultural parent they receive input from, based on the parent's utterances in context. Again, this is a more extreme set-up than what we see in real life, where caregivers and children constantly co-construct useful learning situations through joint attention and other strategies (e.g. Trueswell et al. 2016; Yurovsky 2017). However, we are interested specifically in the space where perspective-inference and language-learning meet. That is, in the cases where the learner has to make inferences about the speaker's intentions based on less direct sources of information (which, as discussed above, typically-developing children are able to do from a relatively young age). This particular model design allows us to

explore how co-development between language and mindreading could play out on a cultural evolutionary timescale. Below, we will briefly discuss the different theoretical views that exist about the role that language played in the cultural evolution of mindreading and vice versa.

3 Theoretical views on the evolution of language and mindreading

Based in part on empirical evidence of linguistic input being important for the development of explicit mindreading (see Sect. 2.1), Heyes and Frith (2014) and Heyes (2018, chapter 7) argue that explicit mindreading is a culturally transmitted skill. That is, that expert mindreaders pass on their mindreading skills by "communicating mental state concepts, and ways of representing those concepts, to novices" (Heyes 2018, p. 168). Further evidence in favour of this hypothesis is provided by a large-scale longitudinal twin study by Hughes et al. (2005), which showed that environmental (rather than genetic) factors explained the majority of variance in children's mindreading abilities. Moreover, Hughes et al. found a strong correlation between mindreading and verbal ability, which were both predicted by the same set of environmental factors. Hughes et al. speculate that these factors may be socio-economic status and siblings, which both affect the availability of conversational input about mental states, and have been shown to be predictive of both language and mindreading development (see de Rosnay and Hughes 2006, for a review).

The potential for developing socio-cognitive skills through social learning is not limited to humans. Russell et al. (2011) showed that great apes who were reared in rich socio-communicative environments (specifically: research projects with the aim of teaching them language) performed no worse on tasks assessing their social cognition (communicative skills and understanding of attentional states and eye-gaze) than on physical cognition tasks. This stood in contrast with the performance of great apes from standard captivity environments, whose social cognition skills were poorer than their physical cognition skills. Furthermore, the performance of the 'enculturated' apes was found to be no different from that of 2.5-year-old children (tested by Herrmann et al. 2007) on the social cognition tasks assessing their production and comprehension of communicative signals, and better than that of the children on the attentional state task. The standard-reared apes in contrast performed worse than the children on the task assessing the production of communicative signals, and no different from the children on the other social cognition tasks. Similarly, Lyn et al. (2010) found that enculturated apes outperformed their standard-reared counterparts in an experiment assessing their comprehension of ostensive pointing and vocalisations by human experimenters.

Importantly, empirical and theoretical work that implicates social learning in the development of mindreading very often emphasises the role of conversation with parents and siblings. Language therefore plays a crucial role in the hypothesis that mindreading is culturally transmitted (Heyes and Frith 2014; Heyes 2018, chapter 7). It is hard to imagine how experts in mindreading would "communicate mental state concepts, and ways of representing those concepts, to novices" (Heyes 2018, p. 168) without having language at their disposal. This stands in stark contrast to theories of language evolution which argue that language *requires* sophisticated mindreading

skills (such as fourth-order metarepresentations and full-blown belief understanding), and that therefore language could not have evolved before these skills were in place (Scott-Phillips 2015a, c; Sperber 2000; Tomasello 2008). Although these accounts do not necessarily dismiss the possibility of some form of co-evolution between language and mindreading (see e.g. Sperber 2000), they do each claim that mindreading skills in *Hominini* must first have evolved to reach a relatively high level of sophistication before language could get off the ground.

Moore (2017b, c) offers a different analysis of language use (specifically, of the cognitive requirements of 'Gricean' or 'ostensive-inferential' communication in the sense of the ability to act with and understand communicative intentions) which allows for a much more gradualist perspective. Moore (2017b) shows that there are many contexts in which such communication does not require the ability to represent beliefs proper (in the sense of a 'propositional attitude' or 'representational relation' which can be false; Apperly and Butterfill 2009; Martin and Santos 2016), but where less cognitively demanding representations, such as a 'registration' or 'awareness' relation (Apperly and Butterfill 2009; Martin and Santos 2016) suffice. In the same vein, Moore (2017b) argues that for simpler forms of Gricean communication, first-order metarepresentations can be sufficient, as opposed to the fourth-order metarepresentation as required according to the definition of ostensive-inferential communication of Sperber and Wilson (1986). This view on the minimal socio-cognitive requirements of Gricean communication, combined with Heyes and Frith's (2014) view that explicit mindreading is culturally transmitted, supports the hypothesis that language and mindreading have co-evolved in lock-step (Malle 2002).

Moore (2017c) argues that although our closest living relatives, the nonhuman great apes, do possess the minimal socio-cognitive requirements for Gricean communication (i.e. the ability to act with and understand communicative intentions, albeit with limited range), they have not been under the relevant ecological pressures to evolve these abilities into language. In contrast, our hominin ancestors became more and more dependent on collaborative foraging over the course of their evolutionary history (Sterelny 2012; Tomasello et al. 2012; Whiten and Erdal 2012). Moore (2017c) argues that this brought about a need to coordinate, which in turn may have given rise to "selection pressures for better social attention and responsiveness, and greater motivation to engage in joint attention and to use existing communication abilities to solve new challenges" (p. 813).

In the remainder of this paper, we develop an agent-based model of language evolution in order to explore whether mindreading and language can co-evolve, which would indicate that full-blown explicit mindreading need not have been in place in order for language evolution to get off the ground. In this model, language learning relies in part on the learner's ability to infer another agent's perspective on the world. We show that, given this very simple model of what mindreading is and how it co-develops with language, a populations' success at inferring each other's perspectives can improve solely through the cultural evolution of a useful language. We further show that such a useful language does not evolve in the absence of any environmental pressure or motivation, but does evolve if (i) agents are under a pressure to be good communicators, *or* (ii) if agents are under a pressure to be good perspective-takers. In the former case, informative lexicons evolve because they allow agents to communicate successfully,

but in the latter case, they evolve because they are instrumental in giving agents insight into each other's minds. Before moving on to present this model, however, we first give a brief review of existing computational models that have explored the role of socio-cognitive abilities in word learning and language evolution.

4 Computational models of word learning and language evolution

Computational modelling can help formalise theories of learning and evolution in order to yield empirical predictions and a better understanding of the dynamics involved (Guest and Martin 2020; Zuidema et al. 2019). Here we are interested in the effect of a developmental interdependence between language and perspective-taking on a cultural evolutionary timescale. That is, if learning a given type of language can improve individuals' perspective-taking, but perspective-taking is in turn relevant in acquiring that language, how does this affect the process of language emergence and evolution in a population? Given such a developmental interdependence, under what circumstances would useful linguistic conventions emerge and evolve? And how would the emergence of such linguistic conventions subsequently influence the populations' success at inferring each others' perspectives? The inspiration for the modelling work presented in this paper was drawn from two types of existing models: models of word learning and models of language evolution. Below we briefly review both types of models, focusing specifically on those that look at the interplay between language and elements of social cognition, such as joint attention.

4.1 Models of word learning

As reviewed in Sect. 2.2, children use their ability to infer a speakers' communicative intentions for word learning. Using a mathematical model, Blythe et al. (2016) showed that cross-situational learning—where the set of candidate meanings is narrowed down based on how often each meaning co-occurs with the novel word across different contexts—can be a powerful mechanism for word learning, even if referential uncertainty is infinite (i.e. when there is an infinite number of candidate meanings). However, this was the case only if the learner had some heuristics for ranking candidate meanings according to their plausibility. Two models of cross-situational word learning by Yu and Ballard (2007) and Frank et al. (2009b)—dealing with finite referential uncertainty—present different ways of achieving such a ranking of candidate meanings using social cognition; through joint attention in the Yu and Ballard model, and through intention-assumption in the Frank et al. model. They each tested their model on its ability to learn a lexicon from the same two videos of naturalistic mother-infant interaction from the CHILDES corpus (MacWhinney 2000).

Yu and Ballard (2007) combined cross-situational learning with social cues in the form of prosody and joint attention (see Paulus and Fikkert 2014; Yu and Smith 2012, for empirical evidence of children using these cues in word learning). As training data, they extracted a speech stream and a meaning stream (consisting of all objects in view when a given word was uttered) from the videos of mother-infant interaction,

and gave the words that received prosodic emphasis and the objects that were in joint attention additional weight in the learning algorithm. Yu and Ballard found that this model performed better on learning a lexicon than (i) the baseline cross-situational learning model, (ii) a model using only prosodic cues, and (iii) a model using only joint attention cues.

Using the same videos from the CHILDES corpus, Frank et al. (2009b) trained and tested a Bayesian learning model which, instead of using overt social cues such as joint attention, assumed an unobserved variable that mediates between the objects present in the physical environment and the words that the speaker utters. This unobservable variable simulates the speaker's referential intention. Frank et al. found that this intention-assumption model outperformed several alternative cross-situational learning models, including the baseline model used by Yu and Ballard (2007) (i.e. without joint attention and prosody added). Frank et al. attribute the high precision of their model (i.e. the fact that it learned relatively few incorrect mappings compared to the other models) to two factors. Firstly, the fact that the model can distinguish between words that are used referentially and words that are used nonreferentially, which allows it to leave words that were used without a consistent referent out of the lexicon. Secondly, the fact that the model considers 'empty' intentions as well as referential intentions, which means that it can disregard utterances that do not refer to any of the present objects. Although this is rather simplified compared to what intention-reading amounts to in real-life vocabulary learning, Frank et al.'s model clearly benefits from its ability to assume a mediating factor between the context and a speaker's utterance: the speaker's referential intention.

Frank et al. (2009a), and Frank and Goodman (2014) applied the *rational speech act* model (Frank and Goodman 2012; Goodman and Frank 2016) to the task of word learning. This is a model of pragmatic reasoning in communication, in which the speaker chooses their utterance by reasoning about a listener, in order to maximise the probability that that listener will interpret the utterance as the speaker's intended meaning (for instance by maximising the informativeness of the signal). The models of Frank et al. (2009a) and Frank and Goodman (2014) are based on the intuition that if a learner assumes that the speaker picks their utterances to be maximally informative, this can help the learner determine the meaning of a novel word. Frank et al. (2009a) and Frank and Goodman (2014) assume that the learner always has full knowledge of the speaker's intended referent, and only has to infer which *feature* of that referent the speaker's utterance is referring to, given the context. Thus, in contrast to the Frank et al. (2009b) model described above, where the learner has to infer *both* the speaker's referential intentions and the lexicon, the learner in these models knows the intended referent and only has to infer the lexicon. In this process, the learner assumes that the speaker maximises the informativeness of their utterance by choosing the most specific word to pick out the intended referent given the context. For example, a learner who observes a speaker using a novel word to refer to a red circle in the context of a blue circle will infer that this novel word means RED rather than CIRCULAR, because otherwise the utterance would be uninformative.

Alongside these models, Frank et al. (2009a) and Frank and Goodman (2014) presented a set of experiments in which participants were put in the same position as either the speaker or the learner agent. They found that the answers of adults

and preschool children closely matched the predictions of the models both in terms of production (what is the best utterance to use given a context and referential intention?) and comprehension (what is the most likely meaning given a particular context, referent and utterance?). Thus, the studies of Frank et al. (2009a) and Frank and Goodman (2014) show that an assumption about the nature of communication (that speakers try to be informative; Grice 1975) can help learners infer word meanings.

In sum, the models described above show that incorporating social cues, intention-assumption, and the assumption that speakers intend to be informative, can each aid word learning. However, these models do not take into account the fact that the skill to read intentions itself needs to develop (Yurovsky and Frank 2017), and that this development may in part depend on linguistic input (as reviewed in Sect. 2.1 above). In Sect. 5 below, we present a model of word learning that does take these developmental considerations into account. Firstly however, we will discuss existing models of the role of joint attention in language evolution.

4.2 Models of language evolution and the role of joint attention

Languages are transmitted culturally: new learners acquire them from more experienced individuals through social learning. This process of cultural transmission is captured by the *iterated learning model*, in which a behaviour is acquired through a process of induction based on observations of that behaviour in another individual who has acquired the behaviour in the same way (Kirby 2001; Kirby et al. 2014). Thus, the behaviour is passed along a transmission chain of individuals, and each new learner (or group of learners) can be thought of as a new generation. Individuals of a new generation arrive at their own internal model of the behaviour by observing the externalised behaviour of one or more individuals from the previous generation, and subsequently externalise the behaviour, thereby providing data for the next generation. The iterated learning model has been widely used to simulate language evolution both in computational models and in laboratory experiments, in order to explore the conditions under which certain kinds of linguistic systems evolve (see Kirby 2017; Kirby et al. 2014, 2015; Smith 2018, for reviews).

To our knowledge, no computational models have been published to date which incorporate perspective-taking (in the sense of inferring an internal, unobservable state of another agent) in language evolution. However, two models have explored the role of joint attention in language emergence and evolution. Joint attention forms a precursor to perspective-taking and mindreading in typically developing children (Charman et al. 2000; Moore and Corkum 1994), and is likely to play a role in their word learning (Trueswell et al. 2016; Yu and Smith 2012). It is also a good predictor of language development in children with autism spectrum disorder (Anderson et al. 2007; Siller and Sigman 2008; Toth et al. 2006). Finally, comparative research has shown that where human children readily engage in and initiate joint attention with others, nonhuman primates do not (Tomasello and Carpenter 2005; Tomonaga et al. 2004).

Kwisthout et al. (2008) modelled three different forms of joint attention, which correspond to sequential stages of its development in children: (i) *checking attention*

(in which the child checks whether her caregiver is attending to the same object), (ii) *following attention* (in which the child allows her attention to be directed to an object by her caregiver), and (iii) *directing attention* (in which the child directs her caregiver's attention to an object) (Carpenter et al. 1998). Kwisthout et al. added these three different types of joint attention to a cross-situational word learning model in order to explore how each type can help reduce referential uncertainty, and how this in turn facilitates language emergence in a group of agents. Kwisthout et al.'s (2008) simulation results showed that the first type of joint attention—checking attention—caused populations to always develop an informative lexicon that yielded maximal communicative success. In contrast, following attention and directing attention by themselves rarely or never (respectively) led to the construction of an optimal lexicon, although they did lead to quicker convergence when combined with checking attention. These results reflect the fact that in this model, referential uncertainty is reduced most strongly by checking attention, followed by following attention, and reduced least strongly by directing attention. This is because in checking attention, speaker and hearer are already attending to the same object. In contrast, the helpfulness of following attention depends on whether another object with the target feature happens to be present in the context, and the helpfulness of directing attention depends on whether the hearer has inferred the target meaning correctly. In sum, the model of Kwisthout et al. demonstrates that joint attention can facilitate language emergence by means of reducing referential uncertainty.

Instead of providing agents with joint attention abilities from the outset and turning these on or off, as Kwisthout et al. (2008) did, Gong and Shuai (2012) developed a model that explores how joint attention might co-evolve with language. Gong and Shuai implemented joint attention as the ability to infer a speaker's communicative intention from nonlinguistic information (i.e. on the basis of an environmental cue rather than lexical knowledge). Gong and Shuai assumed that languages were transmitted culturally, while levels of joint attention (i.e. the probability of correctly inferring communicative intentions from environmental cues) were genetically inherited. Simulation results showed that in order for populations to construct an informative lexicon , they either had to start out with a relatively high level of joint attention, or be subjected to biological selection on communicative success. Thus, the simulation results obtained with this model support an evolutionary scenario similar to that proposed by Scott-Phillips (2015a, c) and Tomasello (2008), where hominins first underwent biological selection for more sophisticated mindreading skills, before language evolution could get off the ground (as discussed in Sect. 3). In Gong and Shuai's model, such biological selection on communicative success caused the level of joint attention to increase over generations, which in turn allowed successful lexicons to be transmitted more faithfully over generations.

5 Incorporating perspective-taking in a model of language evolution

Here, we present a model of iterated language learning in which agents' ability to infer a speaker's referential intention from the context depends on whether they correctly infer the speaker's perspective on the world. This means that when an agent is learning

a lexicon from a cultural parent, their success at doing so depends on how well they infer their parent's perspective (which they have to do simultaneously with learning the lexicon). The developmental dynamics that result from this model, which are described in more detail in Woensdregt et al. (2016), are that a Bayesian learner can solve this joint inference problem by bootstrapping their learning of one attribute of a speaker (the lexicon) from their developing knowledge of the other (the speaker's perspective), and vice versa. In Woensdregt et al. (2016), we showed that this co-development could only get off the ground if the parent's lexicon was at least somewhat informative. This leads to a follow-up question: Under what circumstances can a population of agents who develop in this way *evolve* an informative lexicon from scratch? This is the question we focus on here.

In contrast to Gong and Shuai's (2012) model described above, the model we present here does not include biological adaptation of agents' underlying ability to infer intended referents from non-linguistic cues: every agent enters the population with the same learning abilities. However, lexicons are culturally transmitted over generations (through iterated learning), and can thereby adapt to the (selection) pressures that the population is exposed to. Because agents' perspective-learning benefits from receiving input from a helpful (i.e. informative) lexicon, agents' success at inferring others' perspectives can also change over generations. In other words, perspective-learning can be facilitated by culturally-evolving lexicons. Thus, this model investigates the *cultural* co-evolution of lexicons and perspective-inference. It explores under what circumstances a population of agents can culturally evolve an informative lexicon from scratch, when the faithful transmission of such a lexicon depends on correctly inferring perspectives, and correctly inferring perspectives in turn depends on the lexicon being informative in the first place. The model of language that we use is simple: a lexicon of associations between signals and referents. It therefore cannot capture such linguistic structures as sentential complements or mental state verbs (as discussed in Sect. 2.1). This language model can be made more elaborate in future work however, as we discuss in Sect. 7.

5.1 Model of mental states

As a simple model of how mental states influence communicative behaviour (where we take communicative behaviour to simply be an act of reference), we equate a mental state with a probability distribution over potential referents, from which a speaker's referential intention is then sampled. That is, a speaker's mental state determines how likely they are to choose each of the referents that exist in their world as their intended referent. What gives rise to this mental state is a combination of two factors: (i) the context, which is a state of the world that is observable to all agents, and (ii) the speaker's perspective (i.e. their 'view on the world'), which is not directly observable to other agents. Depending on the combination of context and perspective, each referent will have a particular salience for the speaker, which determines how likely it is to become the speaker's referential intention in that context. This means that if another agent knows what the speaker's perspective is, they would be able to predict how likely the speaker is to talk about each of the possible referents in a given context.

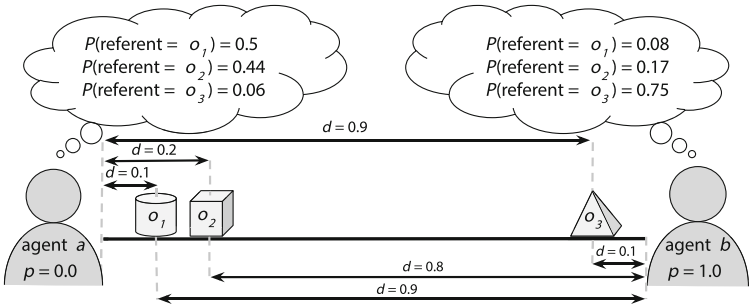

Fig. 1 Diagram of how a context and an agent's perspective together give rise to a probability distribution over potential referents (illustrated using spatial interpretation of perspectives and referent values as described above). Here, the perspectives (p) of the two agents are diametrically opposed; hence their position on opposite extremities of the 'context' line. d stands for the distance between an agent and an object. Thus, object 1 (o_1) is equally close to agent a as object 3 (o_3) is to agent b. However, agent b is 1.5 times more likely to choose o_3 as an intended referent than agent a is to choose o_1 (see the agents' thought bubbles for the probabilities with which they will choose each of the objects as their referential intention). This asymmetry is a result of the distances (and therefore the saliencies) of the other objects in the context. For agent a, o_2 is also very salient, which means the probability mass over potential referents has to be distributed almost equally over o_1 and o_2. For agent b in contrast, the next most salient object (o_2) is much further away than their most salient object (o_3), which means the ratios between the probabilities of choosing objects as intended referents work out differently. This difference in intention probability ratios between the two perspectives is a result of the way mental states are calculated from a set of saliency values (see Eq. 1): the model assumes that saliencies are relative. Therefore, saliency values are normalised to yield a probability distribution over referential intentions

A context in this model consists of a set of values within a one-dimensional range (0.0, 1.0), where there is one value for each referent that exists in the agents' world. We can think of these values as varying along some objective feature dimension that, depending on someone's subjective perspective, can determine a referent's salience. Possible perspectives fall within this same one-dimensional space (but below we constrain them to consist only of the two extremes of that space: 0.0 and 1.0). This allows us to calculate a 'distance' between the agent's perspective and each of the possible referents, and we take the inverse of this distance (i.e. the 'proximity'), normalised over all possible referents, to be the salience of a given referent (see Fig. 1). Equation 1 captures this procedure for how a given perspective and context give rise to a particular referent's (r') probability of becoming the speaker's intended referent ($r_{intended}$).

$$P(r' = r_{intended} \mid p, c) = \frac{1 - |p - c_{r'}|}{\sum_{r \in R} 1 - |p - c_r|} \tag{1}$$

where r stands for referent, p for perspective, c_r for the value of referent r in context c, and R for the full set of referents that exist in the world. An agent's perspective can be interpreted in a literal sense, where the physical proximity of objects influences their salience (as illustrated in Fig. 1), or in an abstract sense: as a world view that determines what potential topic of conversation is most salient to the agent in a given context (e.g. items on a menu for a vegetarian versus a carnivore; paintings in a museum for someone who likes abstract art versus someone who likes realism, etc.).

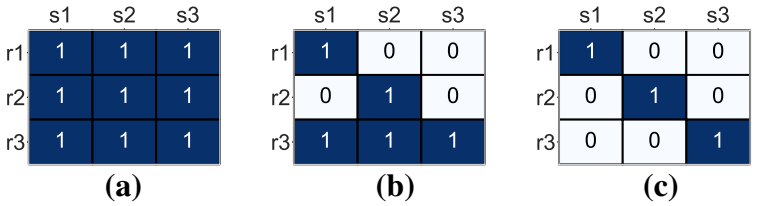

Fig. 2 Examples of lexicons of different informativeness levels (measured as communicative accuracy of lexicon 'with itself', given an error probability ϵ of 0.05). In these matrices, referents are represented by rows and signals by columns. Blue squares (and 1s) represent an association between the corresponding referent and signal, while white squares (and 0s) represent the absence of such an association. **a** Example of a minimally informative lexicon (informativeness = 0.33...), **b** Example of a partially informative lexicon (informativeness = 0.51), **c** Example of a maximally informative lexicon (informativeness = 0.90)

5.2 Model of lexicons

Aside from a perspective, each agent has a lexicon that determines what signal(s) they use for a given referent. Lexicons consist of discrete binary mappings between referents and signals (i.e. a given referent and signal are either associated with each other or not). In order to produce an utterance, a speaker chooses randomly between all signals that are associated with their intended referent, with a small probability (ϵ) of making a production error, which consists of randomly producing one of the signals that are *not* associated with the intended referent. (In all simulations reported below the probability of such errors is set to 0.05.) This procedure for utterance production is captured in Eq. 2, which gives the probability of a signal s being uttered for a referent r given lexicon ℓ.

$$P(s \mid r, \ell) = \begin{cases} \dfrac{1-\epsilon}{|s_r|} & \text{if } s \text{ maps to } r \text{ in } \ell \\ \dfrac{\epsilon}{|\mathcal{S}| - |s_r|} & \text{otherwise} \end{cases} \tag{2}$$

where $|s_r|$ stands for the number of signals that map to referent r, and $|\mathcal{S}|$ for the total number of signals in ℓ. All simulations reported below were run with a world consisting of three possible referents, and lexicons containing three signals.

This lexicon size is of course very small compared to real-world vocabularies, but it is sufficient to capture the situation that we are interested in: one in which learning a lexicon depends in part on learning something about others' perspectives in order to infer their referential intentions. Moreover, three referents combined with three signals are sufficient to create a wide range of lexicon types, varying from completely uninformative to optimal for communication, with a multitude of levels in between. We define the informativeness of a lexicon as its communicative accuracy 'with itself'; that is, the probability that a speaker and hearer who both use that lexicon would understand each other. Using communicative accuracy as a measure of informativeness thus allows us to categorise lexicons according to how much information they provide about the speaker's intended referents. Figure 2 shows examples of a minimally informative, partially informative, and maximally informative lexicon.

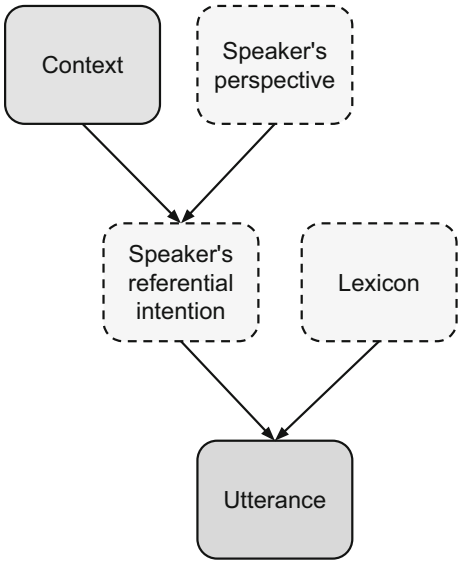

Fig. 3 Diagram of how context, perspective and lexicon together give rise to utterance productions. Variables in dark grey are observable to the learner, variables in light grey are unobservable. The learner has to infer the perspective and lexicon of their cultural parent based on observations of the parent's utterances in context

5.3 Model of learning

The task of the learner in this model is to simultaneously learn the lexicon and perspective of a cultural parent that they receive input from, based on observing what signals the parent utters in different contexts. In other words, the learner has to infer two variables which are unobservable: the parent's perspective and the parent's lexicon, based on two variables that *are* observable: the parent's utterances and the contexts in which they occur. Figure 3 depicts how these unobservable and observable variables are related. The world generates a context[2] which is observable, and that context combined with the speaker's perspective gives rise to the speaker's probability distribution over referents (i.e. their mental state), from which a referential intention is then sampled. That referential intention combined with the speaker's lexicon leads to an utterance, which is observable to the learner.

The model of learning that we use is Bayesian inference (see Perfors et al. 2011, for a tutorial introduction to Bayesian models of cognitive development). Bayesian inference is a model of rational learning by 'hypothesis-testing': a learner considers all possible hypotheses about the system that has generated the data (in this case the speaker, as defined by their perspective and lexicon), and determines how fitting each of these hypotheses is, given (i) the learner's belief in the plausibility of each of the hypotheses prior to having seen any data, and (ii) how likely each of the hypotheses is

[2] In order to speed up simulation run times, the learners in the simulations reported below observed only repetitions of a fixed set of 'maximally informative' contexts. These are contexts that create maximally distinct saliency distributions for the two possible perspectives. Using these contexts does not change anything qualitatively to how learning progresses; the learner simply reaches the same posterior probability distribution more quickly (i.e. with fewer observations), which means simulations complete more quickly.

to have produced the data that the learner observes.[3] In the current model, the learner considers a hypothesis space that consists of all possible combinations of perspective and lexicon hypothesis. This space consists of 343 possible lexicons (i.e. the full space of logically possible lexicons given three referents, three signals, and the constraint that referents need to have at least one signal associated with them), and two possible perspectives, given that we constrain these to the two extremes of the context space. This gives rise to a total of 686 composite hypotheses.

Bayesian learning results in a probability distribution over the full set of composite hypotheses, which is known as the *posterior probability distribution* (because it specifies how much probability the learner assigns to each hypothesis *a posteriori* to seeing the data). In our model, the composite hypothesis consists of both a lexicon hypothesis ℓ and a perspective hypothesis p, so the learner is inferring both these attributes of the speaker simultaneously. The posterior probability of a composite hypothesis given a set of data is proportional to the *likelihood* of the data given the composite hypothesis, multiplied with the *prior probability* of the composite hypothesis, as shown in Eq. 3.

$$P(\ell, p \mid \mathcal{D}) \propto P(\mathcal{D} \mid \ell, p) P(\ell, p) \tag{3}$$

Lexicon hypothesis ℓ and perspective hypothesis p together make up a composite hypothesis about the speaker that the learner can evaluate based on dataset $\mathcal{D}$ (a set of observations of different contexts, each combined with a single utterance produced by the speaker in that context). In order to evaluate the likelihood of a dataset given a particular composite hypothesis, the learner needs to have full knowledge of the generative process; that is, of how a speaker with the perspective and lexicon specified by that composite hypothesis would produce utterances. Therefore, in order to calculate the likelihood of a dataset given a hypothesis, the learner combines knowledge of how a given perspective and context lead to a referential intention (as shown in Eq. 1) and how a referential intention combined with a given lexicon lead to an utterance (as shown in Eq. 2). This process of determining the likelihood of a a dataset $\mathcal{D}$ given a composite hypothesis (ℓ, p) is captured in Eq. 4.

$$P(\mathcal{D} \mid \ell, p) = \prod_{d \in \mathcal{D}} P(s_d \mid \ell, p, c_d) \tag{4}$$

where each individual data point d consists of a context c_d and a signal s_d that was uttered by the speaker in that particular context. The likelihood of a single utterance

[3] Using Bayesian inference to model cognition has been rightfully problematised on the basis that computing the posterior probability distribution over hypotheses easily becomes computationally intractable when the size or complexity of the hypothesis spaces goes beyond that of a toy domain (such as the 3×3 lexicons used here), and that it is as yet unclear how this computational-level process would be implemented in the (resource-bounded) brain (Jones and Love 2011; Kwisthout and van Rooij 2013, 2019; Kwisthout et al. 2011). However, in models of cultural evolution, using Bayesian inference as a model of learning comes with the advantage of full transparency and control over the learning biases that individual learners bring to the task, allowing us to separate these out from the cumulative effects of cultural evolution (Kirby 2017). In addition, using Bayesian inference allows us to build on a large body of existing work on models of iterated Bayesian language learning (see e.g. Griffiths and Kalish 2007; Kirby 2017; Kirby et al. 2007; Thompson et al. 2016).

s_d given a composite hypothesis (ℓ, p) and a context c_d is given by marginalising (i.e. summing) the product of the probability of the signal given a particular referent and the probability of that referent being intended, over all referents in c_d. This process is captured in Eq. 5.

$$P(s_d \mid \ell, p, c_d) = \sum_{r \in c_d} P(s_d \mid r, \ell) P(r = r_{intended} \mid p, c_d) \tag{5}$$

Importantly, all referents that exist in the world are always considered potential referents. That is, each possible referent always has a nonzero probability of being chosen as the speaker's referential intention. This means that in learning the lexicon, the learner cannot exclude any potential signal-referent mapping on the basis of simple cross-situational learning (Siskind 1996). The only way in which contexts differ from one another is in the combination of attributes (i.e values) of the different referents, and observing many different contexts is what gives the learner a way into learning about their parent's perspective.

As mentioned above, the *prior probability distribution* specifies the learner's belief about the plausibility of each of the hypotheses *prior* to seeing the data. In all simulations below, the learner presupposes that all lexicons are equally plausible before seeing any data (i.e. has a uniform prior over lexicons). However, regarding perspectives we compare two different priors: a *uniform prior* which assumes both perspectives are equally likely, and an *egocentric prior* which assigns 0.9 probability to the hypothesis that the cultural parent shares the learner's perspective, and 0.1 probability to the hypothesis that the cultural parent has the opposite perspective. In all simulations presented below, we fix the cultural parent's perspective to be opposite to that of the learner. This means that having an egocentric bias is unhelpful (because it assigns a low prior probability to the hypothesis that corresponds to reality). This egocentric bias is motivated by empirical evidence showing that young children start out reasoning about other minds from an egocentric perspective, and that this bias diminishes over developmental time (see Birch and Bloom 2004, for a review). Over all combinations of lexicon and perspective prior, the prior probability of a composite hypothesis (ℓ, p) is simply the product of the prior probabilities of the ℓ and p hypotheses separately: $P(\ell, p) = P(\ell)P(p)$.

The developmental dynamic that this model results in is described in Woensdregt et al. (2016), and we will summarise it in Sect. 6.1. The addition that we make in the current paper is to embed this model of learning in a population model, which allows us to simulate cultural evolution, as we will describe below in Sects. 5.4 and 5.5.

5.4 Model of cultural transmission

Lexicons are transmitted culturally, across generations of a population, through iterated learning (Kirby 2001). That is, each new agent that enters the population learns their lexicon by induction on data produced by the previous generation. In our simulations, each agent of a new generation receives input (120 <context, utterance> pairs) from a single cultural parent from the previous generation. Importantly, we initialise the very

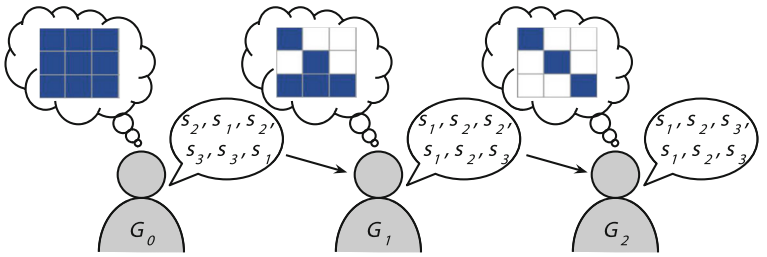

Fig. 4 Diagram of iterated learning. Lexicons are represented as tables with referents on the rows and signals on the columns. Blue squares represent an association between the corresponding referent and signal, and white squares represent the absence of an association. An agent from the very first generation (G_0) produces data based on their uninformative lexicon; this data is then observed by a learner from the next generation (G_1), who induces their own lexicon (and their parent's perspective) based on the data. After lexicon induction, the agent from generation G_1 produces data for generation G_2, and so on

first generation with all agents sharing the same completely uninformative lexicon (one that maps each signal to each referent, as shown in Fig. 2a), so that populations have to start from scratch when evolving informative lexicons. Perspectives are assigned in such a way that they are uniform within a generation but alternate across generations, such that cultural parents always have the opposite perspective to that of their learners. After receiving data from their cultural parent, and updating their posterior probability distribution accordingly, the learner selects a composite hypothesis by sampling from this distribution (Griffiths and Kalish 2007). That is, the probability of a particular hypothesis being selected is equal to the probability assigned to that hypothesis in the learner's posterior distribution. The learner then uses the selected perspective as a model of their parent's perspective, and adopts the selected lexicon as their own. In contrast to the lexicon, the learner's own perspective does not change as a result of learning; what changes is what they believe about their parent's perspective. Figure 4 illustrates how lexicons are passed on over generations through iterated learning.

5.5 Model of selection pressures

All simulation results presented below were obtained with populations of 100 agents, where each new generation is formed by replacing all agents of the previous generation at once. Every new agent receives data from a single cultural parent. We explore the effects of two different selection pressures: *Selection for communication* and *Selection on perspective-inference*. A selection pressure determines how likely a particular agent is to become a cultural parent for agents of the next generation (where a single cultural parent can have multiple learners). This model of selection corresponds to what Boyd and Richerson (1985) termed *natural selection of cultural variants*; where an individual's opportunities of becoming a cultural parent depend on properties of their culturally-acquired trait. We contrast these two selection conditions with a *No selection* condition, where every agent has an equal chance of being selected to transmit data to agents of the next generation.

In the *Selection for communication* condition, the probability that an agent is chosen as a cultural parent is determined by their success at interpreting the utterances of their own cultural parent from the previous generation. In order to interpret an utterance, the agent (referred to below as l for learner/listener) uses Bayes' rule to derive the probability that their cultural parent's referential intention was a particular r, given that they produced signal s (similarly to the listener in the *rational speech act* model; Goodman and Frank 2016). To do so, the listener uses their own lexicon (ℓ_l) and their model of their cultural parent's perspective (p'_{cp}), either or both of which may not correspond to the real lexicon and perspective of the cultural parent.[4] This interpretation procedure is captured in Eq. 6.

$$P_l(r \mid s,\ \ell_l,\ c,\ p'_{cp}) \propto P(s \mid r,\ \ell_l)\, P(r \mid c,\ p'_{cp}) \tag{6}$$

where $P(s \mid r,\ \ell_l)$ is the probability that the cultural parent produces signal s assuming that r is their intended referent and ℓ_l is their lexicon (this probability was specified in Eq. 2 above), and $P(r \mid c,\ p'_{cp})$ is the probability that r is indeed the cultural parent's intended referent, given the context c and the learner's model of their parent's perspective p'_{cp} (this probability was specified in Eq. 1 above).

The communicative success between a cultural parent cp and their learner/listener l in a context c is defined as the total probability that cp will produce a signal which enables l to correctly identify cp's intended referent, over all possible referents, as shown in Eq. 7. (This equation combines the production and interpretation procedures as described in Eqs. 5 and 6 respectively.)

$$cs(cp, l \mid c) = \sum_{r \in R} \sum_{s \in S} P(s \mid r, \ell_{cp}\, p_{cp}, c) \cdot P_l(r \mid s,\ \ell_l,\ c,\ p'_{cp}) \tag{7}$$

where R stands for the full set of potential referents, $\mathcal{S}$ for the full set of signals, ℓ_{cp} for the lexicon of the cultural parent, ℓ_l for the lexicon of the learner, p_{cp} for the parent's real perspective, and p'_{cp} for the learner's model of the parent's perspective (which may be false). For a given agent, we obtain a final measure of communicative success $CS(cp, l)$ by averaging $cs(cp, l \mid c)$ over six randomly generated contexts.

Because these contexts are generated randomly, they are different from the fixed set of maximally informative contexts that each learner is trained on (see footnote 2).

In the *Selection on perspective-inference* condition, an agent's probability of being chosen as a cultural parent is proportional to the posterior probability they assign to the correct hypothesis about their cultural parent's perspective. That is, the better an agent has learned their parent's perspective, the more likely they are to be chosen as a cultural parent themselves.

[4] Because the listener uses their model of their cultural parent's perspective in interpretation, our measure of communicative success takes into account only the learner's comprehension success, not their success at producing signals that their cultural parent will understand, which would require an additional learning phase in which the cultural parent learns about the perspective of the learner. This keeps the *Selection for communication* condition as similar as possible to the *Selection on perspective-inference* condition (described below), in the sense that all that matters for the learner's probability of being selected is the learner's knowledge about their cultural parent, not the cultural parent's knowledge about their learner.

The two selection pressures described above can simulate environmental selection just as well as cultural selection. In the former case (environmental selection), more successful agents are more likely to survive or have more offspring, and are thus more likely to act as 'models' that agents of the next generation learn from. In the latter case (cultural selection), selection is driven by the learners: they choose which agent from the previous generation they learn from, based on that agent's success (Boyd and Richerson 1985). In the model presented here, we do not specify which of these two mechanisms of selection is at play; we only differentiate between the *trait* that is selected on: communicative success or success at inferring perspectives.

Both these pressures could be the result of a need for increased social coordination (Moore 2017c), which (as discussed in Sect. 3) could have arisen when our hominin ancestors became more and more reliant on collaborative foraging (Tomasello et al. 2012; Sterelny 2012; Whiten and Erdal 2012), or as a result of cooperative breeding (Burkart et al. 2009). A pressure on perspective-inference alone (i.e. on the ability to accurately explain and predict others' behaviour by modelling their mind) could perhaps have come from a need for "Machiavellian intelligence", resulting from increased social complexity (Byrne 1996). The other way around, a pressure for communication in the absence of a pressure on perspective-inference is less easy to imagine, because of language's reliance on reading and sharing intentions (which in turn requires a certain level of mindreading) (Tomasello et al. 2005). Even if it is likely that these two selection pressures would have co-occurred during the relevant period in hominin evolution, separating them out in this model is still a useful exercise, because it allows us to look at their effects in isolation.

6 The emergence of an informative lexicon improves perspective-taking over generations

Under what circumstances can a population of agents whose lexicon-learning and perspective-inference are interdependent evolve an informative lexicon from scratch? Below we show that, given the assumptions of our model described above, this does not happen in the absence of any external pressure or motivation on the part of the agents. However, a selection pressure in favour of *either* successful communication *or* successful perspective-inference can be sufficient for maximally informative lexicons to evolve. When this happens, it leads to an increase not just in populations' communicative success, but also in their success at inferring others' perspectives. This is a result of the fact that a more informative lexicon provides agents with better insight into others' perspectives compared to a less informative lexicon. Before we turn to these simulation results on the evolutionary timescale, however, we will first briefly summarise the developmental dynamics that ensue from the model of learning described above.[5] These developmental results are discussed in more detail in Woensdregt et al. (2016).

[5] The code that was used to run all simulations reported in this paper is freely available at https://github.com/marieke-woensdregt/model_coevolution_language_mindreading.

6.1 When perspectives play a role in utterance production, word learning and perspective-taking co-develop

The developmental dynamic that the model of learning described above gives rise to is one of co-development: a Bayesian learner can bootstrap their learning of a speaker's lexicon from their developing knowledge of the speaker's perspective and vice versa. As shown in Woensdregt et al. (2016),[6] Fig. 5 shows that the learner is able to solve their joint inference problem as long as the following two conditions are met: (i) the speaker uses a lexicon that is not completely ambiguous, and (ii) the learner is able to represent the speaker's true perspective.

The more ambiguous the lexicon, the longer the learner takes to correctly infer both lexicon and perspective. However, as long as the lexicon provides at least *some* information (just one signal that isn't associated with all referents is enough), the learner will ultimately get there. The only lexicon type that does not allow the learner to correctly infer *both* the speaker's lexicon and their perspective is the 'minimally informative' lexicon type. This is in fact an *un*informative lexicon type, which consists only of lexicons that associate every signal that they make use of with all referents; these lexicons are therefore completely ambiguous. When receiving input from such a lexicon, the learner does infer the lexicon correctly, but not the speaker's perspective.

What gives learners a 'way in' to their joint inference problem is the fact that they can learn something about the speaker's perspective even before they know the lexicon (i.e. what signals map to which referents), as long as they get to observe a sufficient number of different contexts. This is a result of the fact that the speaker's saliency distribution over possible referents is normalised over referents (because we assume salience is relative). This means that the same combination of a given perspective and a given referent attribute (i.e. the referent's 'position' in the context space) can yield different saliency values, depending on the attributes of the *other* referents. Thus, even if a learner does not know how the speaker's utterances map onto individual referents, they will still be able to evaluate the probability of different perspective hypotheses based on ratio differences between the intention probabilities that those perspective hypotheses predict. However, this learning strategy requires two conditions to be met: (i) that the learner gets to observe a sufficient number of different contexts (with their corresponding utterances), and (ii) that the speaker's lexicon is not entirely ambiguous.

If the learner has a strong *un*helpful bias about the speaker's perspective, as is the case for the egocentric learner (given that we set the speaker's perspective to be opposite to that of the learner), this slows learning down somewhat, but does not prevent the learner from ultimately correctly inferring the speaker's lexicon and perspective (as shown in in Fig. 5b). The only exception to this is the case of the learner who has a prior belief of 0.0 in the correct perspective hypothesis, which is equivalent to a learner who cannot represent the possibility that the speaker might have the perspective opposite to their own (i.e. the "No ToM" learner in Fig. 5).

Learning about the lexicon and learning about the speaker's perspective thus go hand-in-hand. This is not surprising given the design of the model: because the speaker's utterances are a result of an interaction between the speaker's perspective, the

[6] Note that there a smaller lexicon size of two referents and two signals was used.

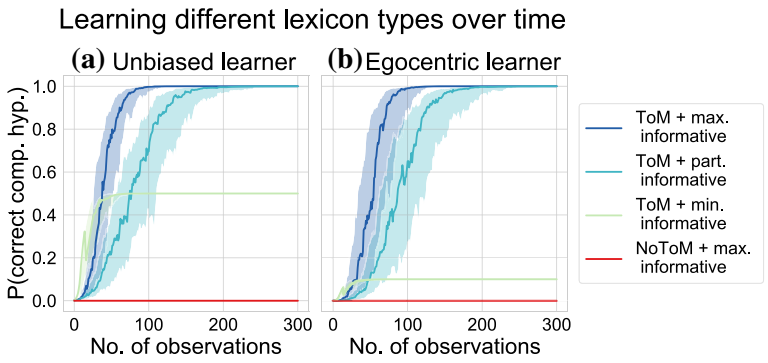

Fig. 5 Time courses of learning for an unbiased learner (**a**) and an egocentric learner (**b**) receiving data from different lexicon types. **a** and **b** each show (i) a learner receiving input from a maximally informative lexicon, (ii) a learner receiving input from a partially informative lexicon, (iii) a learner receiving input from a minimally informative (i.e. *un*informative) lexicon, and (iv) a learner who cannot represent the possibility that the speaker's perspective might be different from their own (hence "No ToM" for "no theory of mind"), who receives input from a maximally informative lexicon. Time courses show the amount of posterior probability assigned to the correct composite hypothesis (i.e. lexicon + perspective) over time (i.e. number of observations). Lines show median and shaded areas show upper and lower quartiles over 100 independent simulation runs per condition

current context, and the speaker's lexicon, a learner who is able to infer the speaker's perspective gains information about which referent the speaker's utterance is likely to refer to. (Note however that even with perfect knowledge of the speaker's perspective, a learner would not be able to predict exactly which referent the speaker will choose in a given context, because this choice is made probabilistically.) Conversely, if the learner knows the lexicon, this provides information about the speaker's referential intentions in different contexts, and therefore about the speaker's perspective.

As mentioned above, this model is a simulation of the hypothesis that language aids the development of mindreading because it provides conversational experience that highlights differences in perspectives (Harris 2005), more than the hypothesis that specific aspects of language, such as labels for mental states (Olson 1988) or sentential complement syntax (de Villiers and Pyers 2002) are necessary to reach full-blown mindreading. As we reviewed in Sect. 2.1, the hypothesis that such conversational experience by itself plays a role in mindreading development was corroborated by the 'discourse only' condition in a training study of Lohmann and Tomasello (2003).

6.2 Without selection, maximally informative lexicons are unlikely to evolve

Turning now to the results of cultural evolution, we first explore what lexicons evolve when populations aren't exposed to any selection pressure, and how this affects populations' success at communicating and inferring perspectives. In this *No selection* condition, the only factors that influence the outcome of iterated learning are (i) the *transmission bottleneck* (i.e. the number of observations each learner gets to see; Kirby 2001), and (ii) the learners' prior probability distribution (which is uniform over lexicons, but, in the case of egocentric learners, not over perspectives).

In all simulations below, we used a transmission bottleneck of 120 observations. At this bottleneck size, an egocentric learner can on average reach a posterior belief in the correct lexicon hypothesis of at least 0.1, and at most 0.95 (expressed in probability), depending on which lexicon type they receive input from. Thus, no lexicon type is learned perfectly (on average) with this bottleneck, while each lexicon type has at least some chance of being learned correctly. The developmental results summarised above show that more informative lexicons generally require fewer observations to be inferred correctly, meaning they will have a higher chance of being transmitted faithfully (i.e. 'passing through' the transmission bottleneck) and therefore will be more stable over multiple generations of transmission. Thus, we expect the bottleneck by itself to exert some pressure in favour of more informative lexicons. However, the informativeness of a lexicon is not the only factor that predicts how long it takes to learn[7], which means that the transmission bottleneck by itself will not solely select for informative lexicons.

As for the effect of the egocentric perspective bias on iterated learning: the learning results summarised above show that egocentric learners take a bit longer to correctly infer the lexicon of their cultural parent than unbiased learners do (how much longer exactly depends on the lexicon). Therefore, a given lexicon should be more likely to be transformed (i.e. altered) from one generation to the next in a population of egocentric learners than in a population of unbiased learners.

Below we show how populations' lexicons evolve under the different selection pressures described above, and how this in turn affects populations' success at communicating and inferring perspectives. In order to visualise how the distribution of lexicons changes over generations, we group all 343 possible lexicons into three main types: (i) minimally informative lexicons, which are in fact *un*informative, (ii) partially informative lexicons, which comprises a large range of informativeness levels, from almost completely ambiguous to almost completely informative, and (iii) maximally informative lexicons. These three main types consist of 7, 330, and 6 lexicons respectively. As mentioned above, we initialise each population with a lexicon of the minimally informative type (specifically, the one that associates every signal with every referent, as shown in Fig. 2a), and we show how the lexicons in the populations change over the course of 400 generations.

Figure 6 shows the results of cultural evolution in the *No selection* condition. The top panel shows that populations in this condition quickly converge on a stable state where partially informative lexicons are strongly dominant. The middle panel shows the distribution of lexicon types present in the populations after convergence, which comes quite close to what would be expected if agents were simply choosing lexicons at random (indicated by the baseline distribution). This baseline distribution simply reflects how many different possible variants there are of each lexicon type (7 minimally, 330 partially, and 6 maximally informative, as mentioned above). Thus, the partially informative lexicons come to dominate in this condition simply because there are so many possible variants of them. We do however see that the maximally infor-

[7] Other factors that determine how quickly a lexicon is learned are: (i) the number of signals it makes use of, (ii) the number of unambiguous mappings it contains, and (iii) whether the learner needs knowledge of the speaker's perspective in order to infer it correctly (which is the case for all lexicons except for those of the 'minimally informative' type).

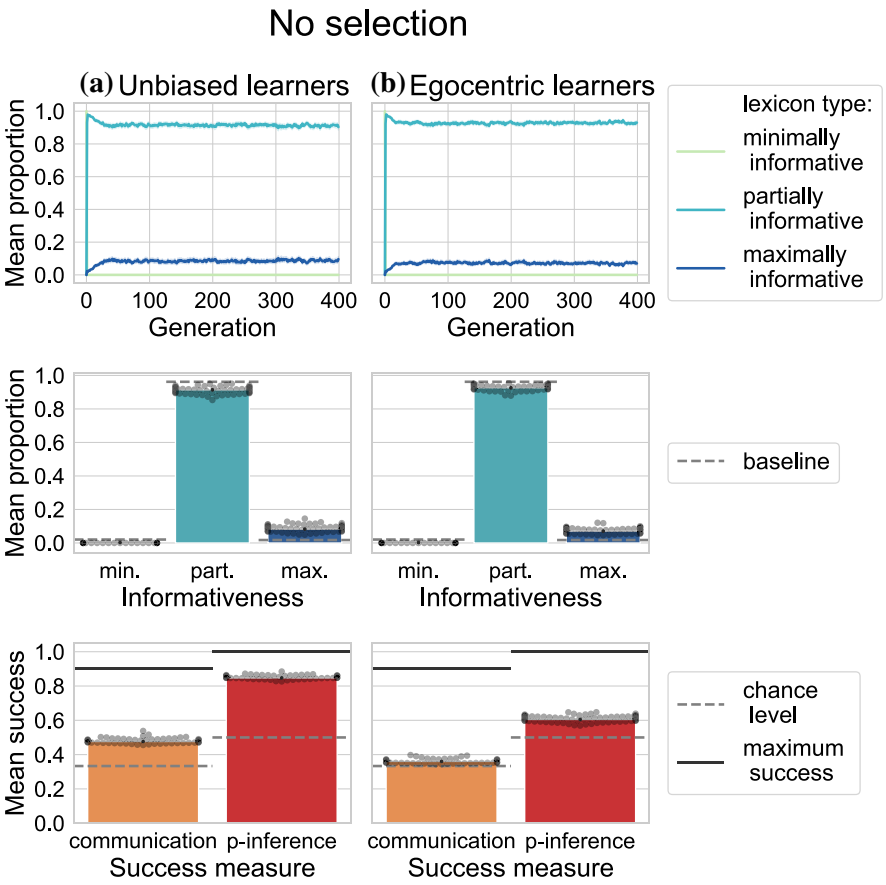

Fig. 6 Time courses of evolution (top), distributions over lexicon types after convergence (middle), and success after convergence (bottom) in the *No selection* condition, with unbiased populations on the left and egocentric on the right. Time courses show mean and 95% confidence intervals over 100 simulation runs (CIs are very small due to large sample size). On distribution plots, points show average over generations 200–400 (i.e. after convergence) for each simulation run, and bars show mean and 95% CIs over all 100 runs. In middle panel, dashed grey lines indicate baseline proportions made up by each lexicon type in full hypothesis space (i.e. expected distribution if agents choose lexicons at random). In bottom panel, dashed grey lines indicate chance level and solid black lines indicate ceiling (for communication this is based on 3 referents and $\epsilon = 0.05$). When populations are not exposed to any selection pressure, they converge on a distribution over lexicons close to the baseline distribution, which is dominated by partially informative lexicons. Consequently, their success at communicating and inferring perspectives after convergence does not greatly exceed chance level, especially for populations of egocentric learners

mative lexicon type is slightly overrepresented compared to the baseline distribution, which is most likely an effect of the transmission bottleneck selecting for lexicons that are transmitted faithfully (as discussed above), for which the maximally informative ones are a strong candidate. The bottom panel of Fig. 6 shows that—as expected given the dominance of partially informative lexicons—populations' success at communicating and inferring perspectives (after convergence) is not very high. This is

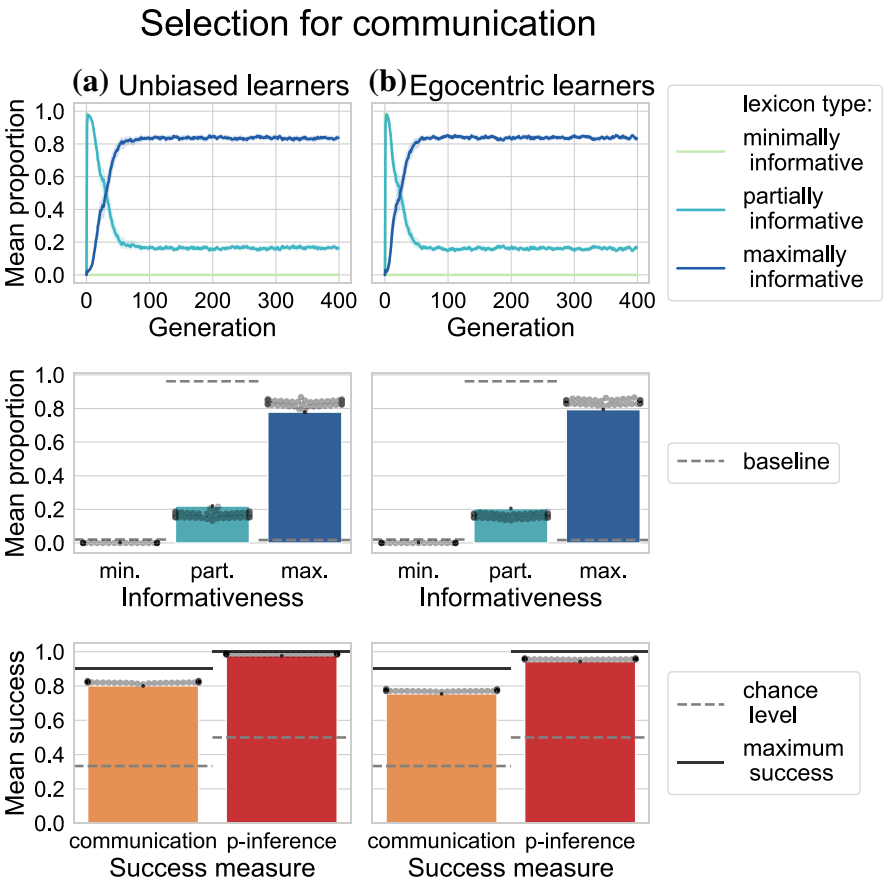

Fig. 7 Time courses of evolution (top), distributions over lexicon types after convergence (middle), and success after convergence (bottom) in the *Selection for communication* condition. When populations are exposed to selection for communication, they overwhelmingly converge on lexicons of the 'maximally informative' type. Consequently, their success at communicating and inferring perspectives after convergence is not far from ceiling, especially in populations of unbiased learners

especially pronounced in populations of egocentric learners, whose success on both measures only just exceeds chance level.

6.3 Selection for communication causes more informative lexicons to evolve, which in turn improves perspective-inference

In the *Selection for communication* condition, the maximally informative lexicon type becomes the majority variant over generations in both unbiased and egocentric populations, as shown in Fig. 7. This is perhaps unsurprising given that this lexicon type provides the most information about a speaker's intended referents, and therefore yields the highest communicative success. The added consequence, however, is that these maximally informative lexicons not only increase populations' communicative

success, but also their success at inferring perspectives. The latter is a result of the fact that the only evidence that learners have available for inferring their cultural parent's perspective is the parent's utterances in context. The more informative the lexicon of the parent (i.e. the more unambiguous, one-to-one mappings it contains) the easier it is for the learner to track the frequencies with which their parent chooses to talk about the different referents in different contexts, and therefore to infer their perspective. Moreover, the more informative lexicons are generally learned more quickly than the less informative ones[8], so a learner who receives input from a maximally informative lexicon will sooner be able to accurately bootstrap their perspective-learning from their lexicon knowledge, compared to a learner who receives input from a partially informative lexicon.

As noted above, there is no biological evolution of agents' ability to learn about perspectives in this model. Within a given population type (i.e. unbiased or egocentric), every agent of every generation enters the population with exactly the same learning abilities and prior probability distribution over hypotheses. The only thing that changes over generations is the lexicons that the agents infer. It is thus the cultural evolution of more informative lexicons that drives the increase in both communicative and perspective-inference success in this condition compared to the *No selection* condition.

6.4 Under selection on perspective-inference, more informative lexicons are more likely to evolve when perspective-inference is hard

Finally, in the *Selection on perspective-inference* condition, we also see that lexicons become more informative over evolutionary time (see Fig. 8), although not to the same extent as in the *Selection for communication* condition. Thus, a pressure for perspective-inference creates selection for more informative lexicons. Just like the finding that a selection pressure for communication leads to an increase in perspective-inference success, this is a result of the fact that the only evidence that learners have available for learning about their parent's perspective is the parent's utterances in context. Although any lexicon that is not completely ambiguous suffices in principle for accurate perspective-inference, the more unambiguous information a lexicon provides about a speaker's referential intentions in different contexts, the quicker the learner can infer the correct perspective. One can compare the situation of a learner receiving input from a very ambiguous lexicon to that of a learner who is trying to divine something about other minds simply through observing others' non-linguistic behaviour. Although this strategy should get an observant learner at least some way towards inferring another's perspective, the data they would rely on to do so would be far sparser and more ambiguous than the evidence received by a learner born into a community that uses a fully-fledged conventional language.

Figure 8 also shows that the increase in maximally informative lexicons over generations is stronger in egocentric populations than in unbiased populations. While in

[8] Note that the number of observations required to learn the different lexicons does not scale perfectly with their informativeness; there are other factors that influence a lexicon's learnability as well. For example, an important exception to this generalisation is that minimally informative lexicons are learned very quickly (on average as quickly as the maximally informative ones), because inferring them correctly does not require any knowledge of the cultural parent's perspective.

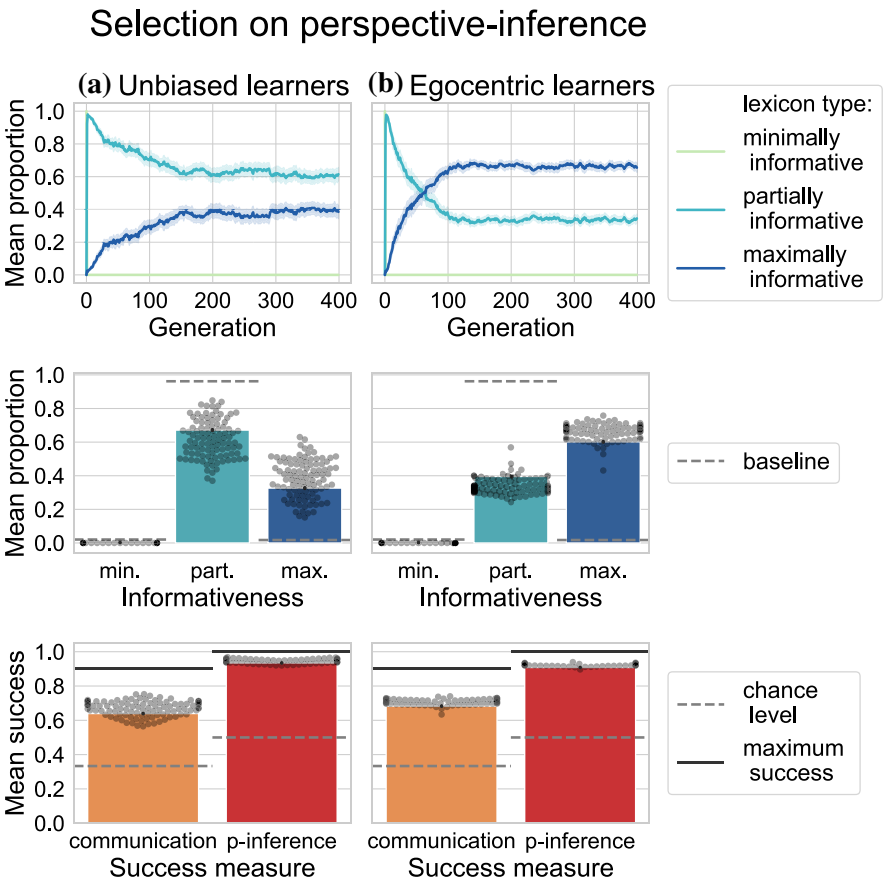

Fig. 8 Time courses of evolution (top), distributions over lexicon types after convergence (middle), and success after convergence (bottom) in the *Selection on perspective-inference* condition. Under selection on perspective-inference, populations tend to converge on a mixture of partially informative and maximally informative lexicons, in a way that allows them to approach ceiling in their perspective-inference success. Populations of unbiased learners select partially informative lexicons more often than maximally informative ones, while populations of egocentric learners show the opposite pattern: they have a preference for maximally informative lexicons

unbiased populations the dominant lexicon type after convergence remains the partially informative one, in egocentric populations the maximally informative lexicons become dominant. This is a result of the fact that egocentric learners have an unhelpful perspective bias, and therefore require more help from the lexicon in order to accurately infer the perspective of their cultural parent. Thus, the (indirect) selection pressure in favour of maximally informative lexicons that results from a (direct) selection pressure on perspective-inference becomes stronger when perspective-inference is hard. This has as a downstream consequence that in the *Selection on perspective-inference* condition, communicative success after convergence is slightly higher in egocentric populations than it is in unbiased populations, which is the opposite effect to what we see in the *No selection* and *Selection for communication* conditions.

7 Discussion

The aim of the current paper was to explore under what circumstances a cultural co-evolution between language and mindreading could get off the ground. Such a two-way positive feedback loop between language and mindreading through cultural evolution could present a middle ground between contradictory theoretical views on whether the emergence of language required sophisticated mindreading, or whether the cultural evolution of sophisticated mindreading required language. Agent-based modelling provides a method for formalising the assumptions that go into such theories and exploring the dynamics that ensue. The co-evolutionary dynamic that we demonstrated here, suggests that a more gradualist scenario of language and mindreading culturally evolving in lock-step is plausible.

In the developmental model described in Woensdregt et al. (2016) , which we built on here, lexicon-learning and perspective-learning co-develop as a result of the simple assumption that a speaker's utterances are not a direct consequence of the context, but rather of an interaction between the context and the speaker's perspective on the world (which is a 'hidden', subjective variable). A Bayesian learner can infer both the perspective and the lexicon of such a speaker after observing the speaker's utterances in a sufficient number of different contexts, by bootstrapping one from the other. However, learning is faster when the speaker's lexicon is more informative, and not possible when the speaker uses an entirely uninformative lexicon, or when the learner lacks the ability to represent the speaker's true perspective. By embedding this model of learning in a population model in which lexicons are transmitted over generations (through iterated learning), we showed that the cultural evolution of informative lexicons leads to agents becoming better at inferring others' perspectives. However, informative lexicons do not evolve under all circumstances. In the absence of any external pressure to communicate or correctly infer perspectives, populations do not converge on maximally informative lexicons, and thus do not become very successful at communicating or inferring perspectives.

In contrast, a selection pressure for *either* successful communication *or* successful perspective-inference can lead to populations converging on maximally informative lexicons. Under a pressure for communication, this happens regardless of agents' prior probability distribution over perspectives, while under a pressure for accurate perspective-inference it happens most strongly when agents have an egocentric perspective bias. The latter is a result of the fact that the more difficult it is for agents to learn a perspective that is different from their own, the more help they need from the lexicon. In general, lexicons that are more optimised for communication provide learners with more information about their cultural parents' referential intentions, which in turn increases the ease with which learners infer their parent's perspective. And conversely, lexicons that are more optimised for accurate perspective-inference also provide learners with more information about their cultural parents' referential intentions, which in turn increases the communicative success of the agents who use those lexicons.

In sum, we showed that the assumption that an agent's perspective-taking development benefits from them receiving input from a useful language, and that perspective-taking in turn facilitates language learning, has consequences on an evo-

lutionary timescale. Firstly, it means that populations of agents can become more successful at inferring each other's perspectives over generations. This happens without the need for anything to change to agents' underlying 'biological' ability to learn about others' perspectives, but instead as a result of the cultural evolution of a language which facilitates perspective-inference. Secondly, we showed that the cultural evolution of a useful language doesn't really get off the ground when there is no external pressure or motivation on the side of the agents to be good at either communicating or inferring each others' perspectives. When such a selection pressure is added, however, either pressure can by itself be sufficient to cause a useful language to evolve, and thus for agents' success at communicating and inferring perspectives to increase over generations.

One may argue that the model of mindreading used here is too simplistic, because its development only consists of the simple task of inferring a single parameter (the 'perspective') which maps from an observable context to a speaker's unobservable saliency distribution over possible referential intentions. Indeed, in its simplest interpretation—that of a speaker who talks more about things that are physically close to them—this model simulates a form of visual perspective-taking that is available to children relatively early, although not at a preverbal age (24 months for understanding *what* another person sees, Moll and Tomasello 2006; and 36 months for understanding *how* things look to another person, Moll and Meltzoff 2011). Level-1 perspective-taking (understanding *what* others see) is also available to non-human primates (Call and Tomasello 2008, Box 1), although level-2 perspective-taking (understanding *how* things look to others) does not seem to be (Karg et al. 2016).

However, as mentioned in Sect. 5.1, our model of what a perspective is and how it influences communicative behaviour is sufficiently abstract to stand in for more complex forms of perspective-taking as well. An agent's perspective simply represents their view on the world, and in combination with a given situation (i.e. context), that worldview will influence how likely they are to talk about different topics. The essential innovation that our model presents (compared to other models of word learning) is that it assumes (i) a mediating factor between the context (be it in terms of the physical situation, the preceding discourse, or the common ground between speaker and hearer) and a speaker's saliency distribution over referential intentions, and (ii) that this mediating factor is not directly observable to other agents, but has to be inferred. This does capture the essence of how mindreading is generally defined (as the ability to attribute mental states to others in order to explain and predict their behaviour), and of how it is taken to play a role in language use (as the ability to infer referential intentions).

In the model presented in this paper, this skill of mindreading by itself is a given: every agent is 'born' with a priori knowledge of how others select referential intentions and utterances. However, this innate knowledge by itself is not sufficient for agents to accurately attribute mental states to others. They first need to infer the other agent's perspective, based on observations of their behaviour (in this case: their utterances in context). The lexicon has an influence on this developmental process: the more informative it is, the easier (i.e. quicker) the process of perspective-inference becomes. It is not impossible for learners to accurately infer another agent's perspective if they receive input from a very ambiguous lexicon; it simply takes longer, and, given a

limited amount of observations, will therefore be less accurate. In other words, the model presented here does not assume that a maximally informative lexicon holds a special key to a 'next level' form of mindreading that cannot be achieved otherwise. We simply assume that receiving input from a useful language is helpful for inferring others' perspectives (and vice versa).

Future work could extend the simple models of mindreading and language used here in several possible directions. Firstly, learners in our model do not have the ability to use what they've learned about the perspective of one agent to inform them when learning about another. For example, if an egocentric learner manages to overcome their egocentric bias by learning from one agent that has a different perspective, they will nevertheless start with the same egocentric bias when learning from another agent. To remedy this, one could extend the model of Bayesian inference of lexicons and perspectives that we used here into a model of *hierarchical* Bayesian learning (Kemp et al. 2007). That would allow learners to infer an 'overhypothesis' about how likely different perspectives are to occur in the population, which would become more accurate the more agents they encounter, and help them infer the perspectives of new agents. Another overhypothesis that learners could infer is the function that maps from the context to the speaker's probability distribution over referents, in which the speaker's perspective is a variable (Eq. 1). As mentioned above, the model presented in this paper assumes that the learner is born knowing this mapping function. However, one could assume that this function needs to be learned as well, thereby adding another dimension to the learner's hypothesis space (mapping function in addition to lexicon and perspective). This would correspond to learners not only inferring another agent's view on the world, but also inferring how having a given view on the world influences an agent's behaviour. Given such an extension, a hierarchical Bayesian learning model would allow the learner to infer (for instance) that the mapping function tends to be the same for all agents they encounter, and use that acquired knowledge when inferring the lexicon and perspective of a new agent.

Secondly, one could imagine a model that includes different kinds of mental states that can be modelled, for instance by distinguishing between percepts, desires and beliefs, or between different orders of recursive mindreading (see de Weerd et al. 2015; Jara-Ettinger 2019; Baker et al. 2017, for existing models of mindreading that could potentially inspire such work). Such a model might result in a developmental staging of when these different aspects of mindreading are acquired.

Thirdly, the model of how lexicon and mindreading interact could be made more complex, such that lexicons can contain certain signals that give the learner more direct insight into another agent's perspective in a way that could not be attained otherwise. This would simulate the hypothesis that certain aspects of language, such as labels for mental states or sentential complement constructions, play a special role in mindreading development (see Olson 1988; and de Villiers and Pyers 2002, respectively). There, our prediction would be that full-blown mindreading could not be reached until the culturally-evolved lexicons start making use of those special signals, which (depending on exactly how communicative success is defined) might happen only under selection for perspective-inference.

Finally, agents in the model presented here do not make as much use of their perspective-taking abilities in communication as they could. Although the

learner/listener does use their model of the speaker's perspective when interpreting utterances, the speaker doesn't make use of this fact when determining which signal to use to convey a given referential intention. The latter could be achieved by adding an extra layer of pragmatic communication on top, for instance using the *rational speech act* model (Goodman and Frank 2016) (see Brochhagen et al. 2018, for a model that combines iterated language learning with cutural transmission of pragmatic communication). Based on the findings of Brochhagen et al. (2018), we predict that adding such pragmatic reasoning skills could change what exactly constitutes a 'useful language' under the two selection pressures we explored here. Certainly under a pressure for successful communication, we would expect that pragmatic agents can make do with somewhat more ambiguous languages. That outcome would in fact more closely resemble what we see in actual natural languages, which show a significant degree of ambiguity (Piantadosi et al. 2012; Wasow et al. 2005) that can be resolved by taking into account the context and reasoning about how one's interlocutor chooses and interprets their utterances.

The evolutionary results that we presented closely match the scenario of language evolution sketched by Moore (2017c) (see Sect. 3). Although agents in all selection conditions share the same innate potential to learn about each others' perspectives and to evolve a maximally informative lexicon, this potential is only realised when populations are under an additional selection pressure to either communicate successfully or infer perspectives successfully. Moore suggests that either (or both) such pressures could have resulted from the need to coordinate that came with an increased dependence on collaborative foraging in our hominin ancestors[9], which is in line with the scenarios of hominin evolution put forward by Tomasello et al. (2012), Sterelny (2012), and Whiten and Erdal (2012). We do not exclude the possibility that such selection pressures have led to small tweaks in the biological underpinnings of sociocognitive skills in the hominin lineage. However, the simulation results we presented here suggest that these selection pressures could each also contribute to the cultural evolution of a useful language, which in turn could lead to better perspective-takers.

The finding that the cultural evolution of a useful language can lead to agents becoming better at perspective-taking, could be taken to be in line with the cultural evolution of mindreading hypothesis put forward by Heyes and Frith (2014). However, there are two ways in which the current model does not fully capture that scenario. Firstly, Heyes and Frith (2014) hypothesise that *explicit* mindreading is culturally transmitted, while they argue that *implicit* mindreading consists of a suite of domain-general neurocognitive mechanisms that are genetically inherited (see also Heyes 2018, Chapter 7). The model of mindreading we used here does not distinguish between implicit and explicit mindreading as such. That is, learners in this model do not learn new ways of representing or reasoning about mental states, as Heyes and Frith (2014) and Heyes (2018, Chapter 7) argue happens in the cultural transmission of mindreading. Instead, the way agents represent and reason about perspectives remains constant across agents' lifespan and across generations, and agents' perspective-inference is simply made easier and more successful when they receive input from a useful lan-

[9] "selection pressures for better social attention and responsiveness, and greater motivation to engage in joint attention and to use existing communication abilities to solve new challenges" (Moore 2017c, p. 813).

guage. Thus, as mentioned above, the developmental model used here comes closest to simulating the hypothesis that language is helpful for mindreading because conversation highlights differences in perspectives (Harris 2005), rather than simulating language as providing a special framework for representing or reasoning about mental states (as hypothesised by Olson 1988; de Villiers and Pyers 2002).

This caveat could be addressed by future extensions of the models of mindreading and language used here, as discussed above. For example, a more complex model of mindreading, which involves representations of different types of mental states (such as desires and beliefs), could be combined with a more complex model of language, which includes a class of signals that are particularly suitable for describing such representations. One could imagine a model where learners can observe their cultural parents describing the (observable) behaviour of other agents, and where languages can evolve to include utterances that describe that behaviour in ways that uncover different types of (unobservable) mental states, which may function better or worse for explaining and predicting others' behaviour. One would expect that the more (complex) mental states the language distinguishes, the better the learners of such a language become at explaining and predicting others' behaviour, which could in turn help them infer others' communicative intentions.

What we presented here, however, is the simplest possible model that allows us to explore the potential for a cultural co-evolutionary dynamic between language and mindreading, in a context that is stripped from as much added complexity as possible. We see this as an important first step towards getting a grip on how language and mindreading may have co-evolved, and how this may have contributed to the cultural evolution of human social cognition.

Acknowledgements We would like to thank Simon Kirby for his invaluable contributions to the design and analysis of the model presented here, as well as to discussions about its theoretical embedding. We would also like to thank two anonymous reviewers, Mark Dingemanse and Riccardo Fusaroli for their helpful comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akhtar, N., & Tomasello, M. (1996). Two-year-olds learn words for absent objects and actions. *British Journal of Developmental Psychology*, *14*(1), 79–93. https://doi.org/10.1111/j.2044-835X.1996.tb00695.x.

Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., et al. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of Consulting and Clinical Psychology*, *75*(4), 594–604. https://doi.org/10.1037/0022-006X.75.4.594.

Apperly, I. (2011). *Mindreaders: The cognitive basis of "Theory of Mind"*. Routledge: Psychology Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–70. https://doi.org/10.1037/a0016923.

Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Astington, J. W., & Baird, J. A. (Eds.). (2005). *Why language matters for theory of mind*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.001.0001.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(March), 1–10. https://doi.org/10.1038/s41562-017-0064.

Baldwin, D. A., & Moses, L. J. (2001). Links between social understanding and early word learning: Challenges to current accounts. *Social Development*, *10*(3), 309–329. https://doi.org/10.1111/1467-9507.00168.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind* (Vol. 74). Cambridge: MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46.

Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8*(6), 255–260. https://doi.org/10.1016/j.tics.2004.04.011.

Blythe, R. A., Smith, A. D., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18–27. https://doi.org/10.1016/j.cognition.2016.02.017.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Brochhagen, T., Franke, M., & van Rooij, R. (2018). Coevolution of lexical meaning and pragmatic use. *Cognitive Science*,. https://doi.org/10.1111/cogs.12681.

Burkart, J. M., Hrdy, S. B., & Van Schaik, C. P. (2009). Cooperative breeding and human cognitive evolution. *Evolutionary Anthropology*, *18*(5), 175–186. https://doi.org/10.1002/evan.20222.

Byrne, D. (1996). Machiavellian intelligence II. *Evolutionary Anthropology*, *5*(5), 172–180. https://doi.org/10.1007/s13398-014-0173-7.2. 1011.1669v3.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192. https://doi.org/10.1016/j.tics.2008.02.010.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4):i–vi, 1–143

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, *15*(4), 481–498. https://doi.org/10.1016/S0885-2014(01)00037-5.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Koshi, B. Webber, & A. I. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. F. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam: North Holland Publishing. https://doi.org/10.1016/S0166-4115(09)60059-5.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7.

de Rosnay, M., & Hughes, C. (2006). Conversation and theory of mind: Do children talk their way to socio-cognitive understanding? *British Journal of Developmental Psychology*, *24*(1), 7–37. https://doi.org/10.1348/026151005X82901.

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*, 1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4.

de Weerd, H., Verbrugge, R., & Verheij, B. (2015). Higher-order theory of mind in the Tacit Communication Game. *Biologically Inspired Cognitive Architectures*, *11*, 10–21. https://doi.org/10.1016/j.bica.2014.11.010.

Eigsti, I. M., de Marchena, A. B., Schuh, J. M., & Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders*, *5*(2), 681–691. https://doi.org/10.1016/j.rasd.2010.09.001.

Frank, M., & Goodman, N. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96. https://doi.org/10.1016/j.cogpsych.2014.08.002.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009a). Informative communication in word production and word learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1228–1233).

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009b). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–85. https://doi.org/10.1111/j.1467-9280.2009.02335.x.

Gagne, D. L., & Coppola, M. (2017). Visible social interactions do not support the development of false belief understanding in the absence of linguistic input: Evidence from deaf adult homesigners. *Frontiers in Psychology*, *8*(June), 1–21. https://doi.org/10.3389/fpsyg.2017.00837.

Gong, T., & Shuai, L. (2012). Modelling the coevolution of joint attention and language. *Proceedings of the Royal Society B*, *279*(1747), 4643–51. https://doi.org/10.1098/rspb.2012.1431.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388.

Grice, H. P. (1975). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 305–315). Cambridge: Harvard University Press.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441–480. https://doi.org/10.1080/15326900701326576.

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. Preprint, PsyArXiv, https://doi.org/10.31234/osf.io/rybh9

Harris, P. L. (2005). Conversation, pretense, and theory of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind*. Oxford: Oxford University Press.

Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *1360*(317), 1360–1366. https://doi.org/10.1126/science.1146282.

Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin and Review*, *22*(2), 313–327. https://doi.org/10.3758/s13423-014-0704-4.

Heyes, C. (2017). Apes Submentalise. *Trends in Cognitive Sciences*, *21*(1), 1–2. https://doi.org/10.1016/j.tics.2016.11.006.

Heyes, C. (2018). *Cognitive gadgets*. Cambridge: Harvard University Press.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*,. https://doi.org/10.1126/science.1243091.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, *76*(2), 356–370. https://doi.org/10.1111/j.1467-8624.2005.00850.x.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. https://doi.org/10.1016/j.cobeha.2019.04.010.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and brain sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2016). Differing views: Can chimpanzees do Level 2 perspective-taking? *Animal Cognition*, *19*(3), 555–564. https://doi.org/10.1007/s10071-016-0956-7.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–21. https://doi.org/10.1111/j.1467-7687.2007.00585.x.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110. https://doi.org/10.1109/4235.918430.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, *24*(1), 118–137. https://doi.org/10.3758/s13423-016-1166-7.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(March), 5241–5245. https://doi.org/10.1073/pnas.0608222104.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/10.1016/j.conb.2014.07.014.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016.

Kwisthout, J., & van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research*, *24*, 2–8. https://doi.org/10.1016/j.cogsys.2012.12.008.

Kwisthout, J., & van Rooij, I. (2019). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior*,. https://doi.org/10.1007/s42113-019-00032-3.

Kwisthout, J., Vogt, P., Haselager, P., & Dijkstra, T. (2008). Joint attention and language evolution. *Connection Science*, *20*(2–3), 155–171. https://doi.org/10.1080/09540090802091958.

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–784. https://doi.org/10.1111/j.1551-6709.2011.01182.x.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, *74*(4), 1130–1144.

Lyn, H., Russell, J. L., & Hopkins, W. D. (2010). The impact of environment on the comprehension of declarative communication in apes. *Psychological Science*, *21*(3), 360–365. https://doi.org/10.1177/0956797610362218.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265–284). Amsterdam: John Benjamins.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005.

Meins, E. (2011). Social relationships and children's understanding of mind: Attachment, internal states, and mind-mindedness. In M. Siegal & L. Surian (Eds.), *Access to language and cognitive development* (pp. 23–43). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199592722.003.0002.

Meristo, M., Hjelmquist, E., & Morgan, G. (2011). How access to language affects theory of mind in deaf children. In M. Siegal & L. Surian (Eds.), *Access to language and cognitive development* (pp. 44–61). Oxford: Oxford University Press.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–46. https://doi.org/10.1111/j.1467-8624.2007.01018.x.

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673. https://doi.org/10.1111/j.1467-8624.2010.01571.x.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613. https://doi.org/10.1348/026151005X55370.

Moore, C., & Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*, *14*(4), 349–372. https://doi.org/10.1006/drev.1994.1014.

Moore, R. (2016). Meaning and ostension in great ape gestural communication. *Animal Cognition*,. https://doi.org/10.1007/s10071-015-0905-x.

Moore, R. (2017a). Convergent minds: Ostension, inference and Grice's third clause. *Interface Focus*, *7*(3), 20160107. https://doi.org/10.1098/rsfs.2016.0107.

Moore, R. (2017b). Gricean communication and cognitive development. *The Philosophical Quarterly*, *67*(267), 303–326. https://doi.org/10.1093/pq/pqw049.

Moore, R. (2017c). Social cognition, Stag Hunts, and the evolution of language. *Biology & Philosophy*, *32*(6), 797–818. https://doi.org/10.1007/s10539-017-9598-7.

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., & Trueswell, J. C. (2009). Use of speaker's Gaze and syntax in verb learning. *Language Learning and Development*, *5*(4), 203–234. https://doi.org/10.1080/15475440903167528.

Olson, D. R. (1988). On the origins of beliefs and other intentional states in children. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 414–426). New York, NY: Cambridge University Press.

Papafragou, A. (2002). Mindreading and verbal communication. *Mind & Language*, *17*(1–2), 55–67.

Parish-Morris, J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development*, *78*(4), 1265–1287. https://doi.org/10.1111/j.1467-8624.2007.01065.x.

Paulus, M., & Fikkert, P. (2014). Conflicting social cues: Fourteen- and 24-month-old infants' Reliance on Gaze and pointing cues in word learning. *Journal of Cognition and Development*, *15*(1), 43–59. https://doi.org/10.1080/15248372.2012.698435.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–21. https://doi.org/10.1016/j.cognition.2010.11.015.

Peterson, C. C., & Siegal, M. (2000). Insights into theory of mind from deafness and Autism. *Mind and Language*, *15*(1), 123–145. https://doi.org/10.1111/1468-0017.00126.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004.

Pyers, J. E., & de Villiers, P. A. (2013). Theory of mind in deaf children: Illuminating the relative roles of language and executive functioning in the development of social cognition. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (3rd ed.). Oxford: Oxford University Press.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, *20*(7), 805–812. https://doi.org/10.1111/j.1467-9280.2009.02377.x.

Russell, J. L., Lyn, H., Schaeffer, J. A., & Hopkins, W. D. (2011). The role of socio-communicative rearing environments in the development of social and physical cognition in apes. *Developmental Science*, *14*(6), 1459–1470. https://doi.org/10.1111/j.1467-7687.2011.01090.x.

Sabbagh, M. A., & Baldwin, D. A. (2005). Understanding the role of communicative intentions in word learning. In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199245635.001.0001.

Scott-Phillips, T. (2015a). *Speaking our minds*. London: Palgrave Macmillan.

Scott-Phillips, T. C. (2015b). Meaning in animal and human communication. *Animal Cognition*,. https://doi.org/10.1007/s10071-015-0845-5.

Scott-Phillips, T. C. (2015c). Nonhuman primate communication, pragmatics, and the origins of language. *Current Anthropology*, *56*(1), 56–80. https://doi.org/10.1086/679674.

Scott-Phillips, T. C. (2016). Meaning in great ape communication: Summarising the debate. *Animal Cognition*, *19*(1), 233–238. https://doi.org/10.1007/s10071-015-0936-3.

Siller, M., & Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental Psychology*, *44*(6), 1691–1704. https://doi.org/10.1037/a0013771.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1–2), 39–91.

Slaughter, V., & Perez-Zapata, D. (2014). Cultural variations in the development of mind reading. *Child Development Perspectives*, *8*(4), 237–241. https://doi.org/10.1111/cdep.12091.

Slaughter, V. P., & Peterson, C. C. (2011). How conversational input shapes theory of mind development in infancy and early childhood. In M. Siegal & L. Surian (Eds.), *Access to language and cognitive development* (pp. 3–22). Oxford: Oxford University Press.

Smith, K. (2018). How culture and biology interact to shape language and the language faculty. *Topics in Cognitive Science*,. https://doi.org/10.1111/tops.12377.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective*. Oxford: Oxford University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (1st ed.). Amsterdam: Blackwell Publishing.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Amsterdam: Blackwell Publishing.
Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B*, *367*, 2141–2151. https://doi.org/10.1098/rstb.2012.0116.
Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 335–364). Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470939345.ch12.
Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *PNAS*, *113*(16), 201523631. https://doi.org/10.1073/pnas.1523631113.
Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–413. https://doi.org/10.1075/prag.10.4.01tom.
Tomasello, M. (2008). *Origins of human communication*. Cambridge: MIT Press. https://doi.org/10.5860/CHOICE.46-3671.
Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, *30*(5), 639–650. https://doi.org/10.1037/0012-1649.30.5.639.
Tomasello, M., & Carpenter, M. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development*, *70*(1), vii–132. https://doi.org/10.1111/j.1540-5834.2005.00324.x.
Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, *23*(1), 157–176. https://doi.org/10.1017/S0305000900010138.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*, 675–735. https://doi.org/10.1017/S0140525X05000129.
Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207.
Tomonaga, M., Tanaka, M., Matsuzawa, T., Myowa-Yamakoshi, M., Kosugi, D., Mizuno, Y., et al. (2004). Development of social cognition in infant chimpanzees (Pan troglodytes): Face recognition, smiling, gaze, and the lack of triadic interactions1. *Japanese Psychological Research*, *46*(3), 227–235. https://doi.org/10.1111/j.1468-5584.2004.00254.x.
Toth, K., Munson, J., Meltzoff, N. A., & Dawson, G. (2006). Early predictors of communication development in young children with autism spectrum disorder: Joint attention, imitation, and toy play. *Journal of Autism and Developmental Disorders*, *36*(8), 993–1005. https://doi.org/10.1007/s10803-006-0137-7.
Townsend, S. W., Koski, S. E., Byrne, R. W., Slocombe, K. E., Bickel, B., Boeckle, M., et al. (2017). Exorcising Grice's ghost: An empirical approach to studying intentional communication in animals. *Biological Reviews*, *92*(3), 1427–1433. https://doi.org/10.1111/brv.12289.
Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 1–71. https://doi.org/10.1016/j.cognition.2015.11.002.
Ünal, E., & Papafragou, A. (2016). Production-comprehension asymmetries and the acquisition of evidential morphology. *Journal of Memory and Language*, *89*, 179–199. https://doi.org/10.1016/j.jml.2015.12.001.
Wasow, T., Perfors, A., & Beaver, D. I. (2005). The puzzle of ambiguity.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x.
Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development*, *82*(3), 780–92. https://doi.org/10.1111/j.1467-8624.2011.01583.x.
Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2119–2129. https://doi.org/10.1098/rstb.2012.0114.
Woensdregt, M. S., Kirby, S., Cummins, C., & Smith, K. (2016). Modelling the co-development of word learning and perspective-taking. In *Proceedings of the 38th annual meeting of the cognitive science society*.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016.

Yurovsky, D. (2017). A communicative approach to early word learning. *New Ideas in Psychology*, *50*, 73–79. https://doi.org/10.1016/j.newideapsych.2017.09.001.

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, *20*(2), 1–17. https://doi.org/10.1111/desc.12349.

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., et al. (2019). Five ways in which computational modeling can help advance cognitive science: Lessons from artificial grammar learning. *Topics in Cognitive Science*,. https://doi.org/10.1111/tops.12474.