Jennifer Nagel, University of Toronto
Forthcoming in: Themes from Williamson, Alex Burri and Michael Frauchiger, eds.
November 1, 2024

Common knowledge and its limits

ABSTRACT: What is common knowledge? According to the dominant iterative model, a group of people commonly knows that *p* if and only if they each individually know that *p,* and they furthermore each know that they each know that *p,* and so on to infinity. According to the integrative model proposed in this paper, a group commonly knows that *p* when its members are united in a state of mind of the type whose contents must be true. Epistemic integration within a group is enabled by symmetrical signalling processes such as eye contact. In conversational dyads, symmetrical processing operates on pairs of signals produced by the two sides in a familiar format: speakers generate content for joint attention in main channel communication, and addressees evaluate that content in backchannel communication. Processes of reinforcement learning shape our pairwise signalling, driving the accumulation of common knowledge, both in response to extrinsic reward for coordinated action, and in response to the intrinsic reward of curiosity. Where the iterative model caps the epistemic performance of the group at the level of its weakest member, the integrative model of common knowledge shows how groups working together can outperform their strongest member working alone.

1. An attractive epistemic power

Human beings have a distinctive capacity to join each other in knowing something. Joint or common knowledge is not simply a matter of happening to know something that is also known by someone else. Two people might each know a fact without knowing it together: for example, they might each discover a secret independently, without ever realizing that the other also knows it. By contrast, in states of common knowledge the key fact is out in the open between the parties involved, paradigmatically by being explicitly recognized in face-to-face conversation, or by being jointly perceived, say by friends who notice something together while out on a walk. However it is gained, common knowledge has a certain natural appeal: across cultures, humans avidly point things out to each other, starting early in childhood (Liszkowski, Brown, Callaghan, Takada, & De Vos, 2012). As adults, we spend a substantial portion of our waking hours immersed in free-ranging conversation, often with no obvious benefits beyond the expansion of what is commonly known (Dessalles, 2020). But if some contrast between private and shared knowledge is easily sensed, this is not to say that it is easily explained: philosophers are currently far from consensus on the nature of common knowledge, let alone the question of why we find it so attractive.

Common knowledge seems to be a problem both for epistemologists and for philosophers of mind. Epistemologists may wonder what can be known in an instance of common knowledge, if anything, and how knowledge of this sort could be structured; philosophers of mind may ask what it is about the human mind that enables two or more individual humans to enjoy a shared state of knowing. Those who follow Timothy Williamson in situating epistemology within the philosophy of mind will want to treat these questions in an integrated fashion. This approach opposes the twentieth-century program of analysis, in which epistemologists attempted to find conditions specifying the type of true belief that constitutes knowledge, with belief seen as

an independent component to be analyzed by a non-epistemic philosophy of mind. That program failed to generate any satisfactory analyses, for reasons Williamson traces to the fundamental character of knowledge. He argues for the recognition of knowledge itself as a mental state; in his view, knowledge is the most general factive mental state, distinguished from other mental states by its necessary restriction to true contents (Williamson, 2000, ch.1). If we consider mental states in terms of how they serve agents embedded in a larger environment, then representational states that must reflect this environment have a certain natural primacy, he maintains, with states of mere belief best understood as derivative byproducts of our vital capacity for knowledge. To start as the program of analysis did, anchoring philosophy of mind in representational states whose content may diverge arbitrarily from reality, is to miss the basic point of representation, obscuring our understanding of intelligent agency. "If we try to leave epistemology out of the philosophy of mind," Williamson contends, "we arrive at a radically impoverished conception of the nature of mind" (2000, 41).

Common knowledge marks a point at which the deliberate incorporation of epistemology into the philosophy of mind stands to enrich our understanding of the mental. Some philosophers have suggested that whatever happens mentally in common knowledge has no importantly epistemic dimension: in their view, attitudes such as common belief or common acceptance pose just the same problems as common knowledge, at least as far as the powers of the mind are concerned (e.g. Heal, 1978, 116; D. Lewis, 1978, 44 fn.13). Meanwhile, others have attempted to treat common knowledge as a theoretical problem to be tackled in abstraction from its realization in minds like ours (e.g. Aumann, 1976). Williamson himself has criticized various resultant theories of common knowledge while offering no positive account of his own, beyond a few suggestive remarks that leave important questions unanswered, notably questions about what could naturally drive us to build states of shared knowledge. The approach to be pursued here follows Williamson in taking knowledge to be the most general factive mental state. From that starting point, the aim is to develop a psychologically realistic model of the phenomenon, a model that situates the pursuit of common knowledge within broader patterns of human learning and motivation.

Section 2 examines the dominant approach to common knowledge, in which the shared state is analyzed as an iterative configuration of the knowledge of the individual participants. This approach is incompatible with some of Williamson's core ideas about knowledge, prompting some philosophers to reject those ideas, and prompting others to reject the existence of common knowledge. Psychological plausibility is a further problem for the dominant approach, and arguably a problem even for variants of the approach that truncate or simplify its complexities.

Section 3 lays out a rival theory of common knowledge, in which the group integrates the epistemic powers of the participating agents. To gain shared knowledge that *p* is to unite with others in knowing that *p*, as opposed to making parallel moves as individuals. The process here is somewhat analogous to multimodal sensory integration within a single agent, where multiple sensory channels can together yield sharper guidance than the best individual channel by itself. When humans interact in a shared environment, a relevantly similar unification can emerge, enabling the group as a whole to outperform its best members as individuals, supporting their coordinated action, and enabling individuals to emerge epistemically enriched from their teamwork.

Focusing on the special case of the conversational dyad, Section 4 describes the architecture of dividing epistemic labor between the complementary roles of speaker and addressee: the speaker always selects content for joint attention, and the addressee always settles whether the dyad accepts that content, for better and for worse. This model is supported by extensive empirical evidence that addressees are not simply passive recipients but active evaluators, producing a distinctive type of 'backchannel' communication that registers in the individual awareness of conversational participants without itself passing into joint awareness. While researchers have done much to expose the surprising scope and extent of backchannel communication, its specifically epistemic value has largely been ignored. Indeed, some have taken the fact that dyads act both on their better and on their worse decisions as a reason to exclude properly epistemic considerations from their models of dyadic interaction. What I will argue is that participants learn from their better and worse decisions in an important way: because the epistemic quality of dyadic interaction tends to make subsequent dyadic behavior either more or less rewarding, ordinary processes of reinforcement learning will lead conversational participants towards signalling in ways that optimize the epistemic quality of their interactions. This optimization stabilizes dyadic interaction, making conversation advantageous for participants on all sides, with common knowledge as a basin of attraction. Excluding epistemic considerations leaves us with a radically impoverished model of conversational interaction; incorporating those considerations enables us to explain otherwise mysterious patterns of human behavior.

A final section examines Williamson's suggestion that we ordinarily treat common knowledge as a default condition, both in light of the dynamics of conversation, and in light of research on the formation of personal and cultural common ground. His suggestion generates a puzzle about what drives us to expand common knowledge, but this puzzle can be solved with the help of materials from his larger epistemic framework.

2. Iterative models of common knowledge

The 'classical' or iterative characterization of common knowledge takes the knowledge of individuals as its starting point, aiming to construct the openly shared state out of a series of iterations of individual knowledge. According to this characterization, a group of people commonly knows that *p* if and only if each member of that group knows that *p* (this is 'level one mutual knowledge'), and furthermore each member of that group knows that each member of that group knows that *p* (level two mutual knowledge), and so on to infinity (for reviews, see Greco, 2015; Vanderschraaf & Sillari, 2022).[1]

Many arguments have been advanced in support of the iterative approach. Notably, philosophers have devised hypothetical scenarios highlighting the gap between common knowledge and any merely finite level of individual mutual knowledge. We expose a difference between the mental state of common knowledge and the mental state of level two mutual knowledge, for example, if we can show that some actions are rational when the parties have common knowledge that *p*, but irrational if each of the parties merely knows that each of them knows that *p*. In a scenario of the type used to support this point (inspired by Heal, 1978), two individual prisoners, let us call them Alice and Bob, are housed separately in the women's and men's wings of the same oppressive prison. This prison has a strange security vulnerability: all locks and alarms are deactivated, leaving prisoners free to escape, if buttons X and Y are both pushed when the bell rings at noon on Sunday, as it does every week. Button X is located on a control panel accessible to women prisoners in their recreation yard; button Y is located on an equivalent panel in the men's yard. However, if only one of these buttons is pressed, there are disastrous consequences for the lone button-presser.

Suppose that each of Alice and Bob learns of the vulnerability, henceforth to be abbreviated as *p*, by finding and studying a security blueprint in an obscure nook of the central prison library, a library which is used by the women and men prisoners at different times of day. Access to synchronous communication, say through contraband smartphones, would make it easy for Alice and Bob to coordinate an escape by pressing buttons X and Y when the bell rings. Merely finite levels of individual mutual knowledge do not seem to support rational coordination in the same way. This is clearest at the first step, where each of Alice and Bob in isolation

[1] A note on terminology: it is a sign of the dominance of this way of thinking of common knowledge that some authors (e.g. Lederman 2018) stipulate that the expression 'common knowledge' is to be understood as equivalent to this iterative characterization. Such authors use terms such as 'public information' for our starting points of pre-theoretical interest: in their terms, it is a live question whether public information is a matter of having common knowledge. For some, it is also a live question whether there really is any such thing as public information – as we shall see, Lederman is ultimately skeptical that there is a natural kind here. As a non-skeptic, I will use 'common knowledge' to pick out the pre-theoretically interesting state shared in conversation and joint attention, but I will argue against the iterative characterization of it.

knows that *p*, but neither knows that the other knows. Because the value of pressing button X (or Y) depends on one's counterpart doing their thing as well, both will refrain from pressing, given that neither has evidence that anyone on the other side has reason to act. But even when Alice learns that Bob knows that *p,* perhaps by being shown hidden camera footage of him studying the diagram in the library, she should still hesitate to push button X when the bell rings: it seems she should now worry that he won't know that she knows, and therefore she should not expect him to press button Y, making it unwise for her to press X. Symmetrical considerations apply on Bob's side: if he is shown a video of Alice in the library, coming to know that *p,* he should have parallel worries about her. The level two mutual knowledge that the parties now have should still leave Bob worrying that Alice will not know that he knows she knows that *p*, and therefore fearing that she will not press button X when the bell rings. We can go on to articulate a further series of steps in which each of these separated characters accumulates more levels of knowledge of the other's epistemic position, including their past exposure to what the other has learned about their learning: Alice can be shown footage of Bob watching footage of her making the initial discovery, and so on. However, it can be argued that similarly paralyzing worries should continue to apply for any finite level of mutual knowledge that these two individuals possess, as long as they are only ever intermittently witnessing each other gaining individual knowledge, and never interacting as a live dyad. If no particular finite level of individual knowledge is satisfactory for rational coordination, never quite matching the openness of synchronous communication, then, at least according to advocates of the classical conception of common knowledge, what happens in the dyad can only be modelled by continuing to infinity.

Of course, everyone agrees that communicators aren't literally or explicitly going through an infinite hierarchy of thoughts whenever something is out in the open between them; in variants of the classical approach, the higher iterations are only potential (D. Lewis, 1969), or we discount and expect each other to drop the more elaborate steps of the hierarchy because we all recognize their cost (Binmore & Samuelson, 2001). However, these idealized variants remain open to multiple lines of criticism, including criticism about their precarity and psychological realism. For example, if communication involves even low levels of recursive theory of mind, it should be taxing (P. A. Lewis, Birch, Hall, & Dunbar, 2017), but the paradigm 'out-in-the-open' situation, spontaneous conversation among friends, feels effortless and inviting.

Timothy Williamson identifies a deeper structural cause for concern with infinitely iterative approaches to common knowledge. He argues that knowledge demands safety from error: "If one believes *p* truly in a case α, one must avoid false belief in other cases sufficiently similar to α in order to count as reliable enough to know *p* in α." (Williamson, 2000, 100). For example, when a subject S knows by looking at an object that it

is more than one meter tall, it must be the case not only that this object is indeed more than one meter tall, but also that if its height were minutely different, S's visually-grounded judgment about whether it is more than one meter tall would not be wrong. To classify an object X as more than one meter tall when one would misclassify an extremely similar object Y is just to be luckily correct about X; such a judgment is too risky to constitute knowledge. Given that human ways of judging qualities such as height have natural thresholds of discrimination, knowledgeable judgments must stay well within these thresholds. The argument is generalized to knowledge of any property that can be gained or lost through gradual change, including knowledge itself, for example as a perceptual signal becomes gradually fainter with distance. Chapter 5 of *Knowledge and its Limits* extends this reasoning to mount an attack on the KK principle, which states that for any pertinent proposition *p,* if one knows that *p,* then one can always know that one knows that *p.* A quick way of condensing the force of this argument runs as follows: if the KK principle were correct, then even in borderline cases of knowledge, one must be able to know that one knows; however, the margin of error constraint on knowledge entails that it must be true that one knows that *p* in any case close to a case of knowing that one knows that *p*, a result which generates a contradiction, given the fact that borderline cases by definition lie close to cases in which knowledge fails to obtain. The idea is not that one can never achieve the iterated state of knowing that one knows, but just that iteration will be available only for relatively central states of knowledge, states robust enough to be really quite dissimilar from cases in which knowledge fails. The margin for error constraint reflects the stringency of knowledge, and greater stringency is required to achieve higher-order levels of knowledge. Meanwhile, troubles with the single-agent intrasubjective case spill over quickly to the intersubjective case involving groups who have higher-order mutual knowledge, because given the factivity of knowledge, there is only a short step from knowing that someone else knows that you know that *p* to knowing that you know that *p*. In Williamson's view, "every iteration of knowledge, intrasubjective or intersubjective, adds a new layer of difficulty" (Williamson, 2000, 134). Accordingly, margin for error principles impose firm limits on the number of iterations of knowledge attainable by limited agents such as ourselves, making infinitely iterative common knowledge humanly impossible.

Some philosophers see this result as problematic either for Williamson's margin for error principles, or for his suggestion that knowledge is a state that can be gained or lost through gradual change, with inherently tricky borderline cases. Daniel Greco, who champions the KK principle and accepts the iterative model of common knowledge, draws attention to the popularity of this model in disciplines beyond philosophy. He notes that "standard arguments against KK typically have the consequence that common knowledge is unattainable. If these arguments are sound, we must reject explanations of linguistic, economic, and other social phenomena that appeal to common knowledge" (Greco, 2014, 170). Defenders of Williamson might

reply that a sound argument against KK undermines any attempted explanation of social phenomena in terms of iterative common knowledge, while perhaps also pointing to evidence that these attempted explanations seem to be empirically inadequate (e.g. as surveyed in Crawford, 2019). However, the sheer scale of the interdisciplinary conflict here could be enough to generate some qualms about which side may have gone wrong.

Others who are more sympathetic to Williamson are skeptical about common knowledge, arguing that it is problematic even if the KK principle is not directly challenged. Harvey Lederman constructs an argument against the possibility of iterative common knowledge, drawing mainly on the idea that layers of difficulty are added at the intersubjective steps, where each party must reason about the epistemic position of the other (Lederman, 2018). This reasoning is driven at each such step by considerations about the slight natural imprecision of judgments extracted from experience, at least for subjects like us. From one's own perceptual impressions, one never gains infinitely precise knowledge of a feature such as the height of an object, for example, but can at best know that this height lies within some range. To evaluate what others might know from this object's appearances, one must expand that range further, to allow for the fact that the object itself might for all one knows have an actual height either towards the top or the bottom end of the range one knows it to have, a range of possible heights whose presumed appearances to the other party will therefore spread over a still larger range. These slight allowances for imprecision accumulate as we move to higher and higher levels of mutual knowledge, with the strength of what is mutually known at higher levels gradually diminishing to zero. A further argument generalizes this result from parameters such as height to contents of any sort: Lederman argues that extended iterations leave us with no very-high-order grasp of the fact that we are even interacting with other agents capable of knowing anything. If common knowledge must be infinitely iterative, Lederman argues, then nothing substantive can be commonly known.

As Lederman sees it, advocates of the iterative conception of common knowledge have gone wrong from the start in assuming that there is a robust pre-theoretical phenomenon here that needs explanation. These advocates set an apparent target (he calls it "public information") when they launch their articles on common knowledge by "introducing readers to a putative natural class of examples where people have public information" and then immediately "pose the question of what psychological features these examples share" (Lederman, 2018, 1090). In his view, "we should reject the presupposition of this question, namely, that there are some relevant psychological features which unite the examples of public information. (…) There is little prima facie reason to think that there is one particular pattern of attitudes which people exhibit in situations as different as listening to a conversation and looking at their surroundings on a casual stroll" (Lederman,

2018, 1090). He suggests that these motivating examples are "so limited that it is not even clear how to identify new examples of the phenomenon, or how to classify even mildly 'hard' cases, as would be needed to assess a conceptual analysis of the notion" (Lederman, 2018, 1090-1).

Even if we are fully convinced by Lederman's arguments against iterative accounts of common knowledge, we might worry that fragmenting public information into multiple disparate phenomena leaves us with the larger burden of explaining those phenomena in a piecemeal fashion. Perhaps he is moving too swiftly to dismiss the existence of a natural class here; indeed, his final gloss on the paradigmatic examples suggests that he may be missing something important that unites them. Merely "listening to a conversation" is not enough for common knowledge: if I surreptitiously eavesdrop on a conversation in which two people are discussing a secret, this secret is not thereby out in the open between the three of us. Building common knowledge in conversation seems to depend on being a recognized participant. Likewise, two people cannot achieve joint attention just by "looking at their surroundings on a casual stroll", if they are oblivious to each other's presence or reactions to the environment. Common knowledge through joint attention involves not just looking, but something more interactive in character. This character generalizes across perceptual modalities: for example, when I press an object into the palm of your hand, under the table, then our active haptic contact supports common knowledge of the existence and location of that unseen object. Attention to the specific epistemic vitality of synchronous interaction can guide the construction of a model that will help with harder cases.

3. The integrative model of common knowledge

In his discussion of the failure of successive efforts to give a reductive analysis of knowledge in terms of true belief and other factors, Williamson considers the objection that this project might gradually edge towards success if we allow it infinite iterations of increasing complexity. At each level, proposed analyses will be rejected, but more complex refinements can be proposed to better fit the set of intuitive cases so far proposed. Williamson's response to that objection suggests a strategy of a type applicable to the present problem: "We can approximate a circle as closely as we like with sufficiently many sufficiently small triangles; it does not follow that we should think of the circle as made up out of triangles. The possibility of approximating knowledge in terms of belief and other concepts is not good evidence for the conceptual priority of belief over knowledge" (Williamson, 2000, 4). In a similar spirit, we can grant that advocates of the iterative approach are right that no finite level of mutual knowledge suffices for common knowledge, just as participants in the program of analysis were right to reject each proposed reductive analysis of knowledge as true belief plus some non-epistemic factors. Where advocates of the iterative conception went wrong was in concluding that common knowledge must therefore consist in (or be ideally approximated by) infinite levels of the mutual

knowledge of individuals. If we are having trouble building common knowledge out of many small triangles of increasingly complex individual knowledge, perhaps we have made a fundamental mistake about what has priority.

In the model I will propose, a group that commonly knows something has a mind of its own, attaining knowledge in a way that takes priority over the members' individual knowledge: the group is a unified epistemic subject in its own right. To develop this model, we will start by considering a less controversial case of epistemic integration. Multimodal sensory integration is psychologically better understood than common knowledge, and it has some useful points of structural similarity.

A typical experiment investigating sensory integration might ask experimental subjects to evaluate which of two metal bars is wider, using either sight (hands-off), touch (eyes masked), or both modalities together. For most people, sight has a lower margin of error than touch on such a task, affording a correspondingly greater range of knowledge. Pairs of bars that are close in width will feel roughly the same to the hand (eyes masked), generating random or inaccurate judgments, even when the eye could see which of those bars is wider, judging safely in the sense that visually-based judgments on similar pairs of bars even slightly closer in width would still be accurate. However, subjects who use both sight and touch together will outperform subjects who use sight alone. Some pairs of bars indistinguishable in width to the eye can be known to be different by the agent who simultaneously sees and touches them. Indeed, the nervous system combines information from these two sensory channels in a statistically optimal fashion, appropriately decreasing reliance on either sensory channel as noise is added to it (Ernst & Banks, 2002). But it is a good question how the unified agent is structured to perform so well, as opposed to, say, reacting in a way that is capped at the performance of the better sense, or worse, dragged down by the less responsive modality.

In one leading model of the relevant epistemic architecture, the integration is achieved by simply adding the neural activity of the sensory channels together for a given stimulus object (Ma, Beck, Latham, & Pouget, 2006). This neural activity can be more or less informative: when making a judgment of a continuous parameter such as width, for example, sensation does not deliver an infinitely precise verdict. Seeing a metal bar that is, say, 70mm wide activates a significant population of neurons encoding various potential sizes for the target object, with some of these neurons firing much more actively than others. Each neuron's firing rate can be represented by a tuning curve showing its response to a range of stimuli—for example, objects of different widths—with the peak of this curve marking the neuron's 'preferred stimulus.' Internal noise plays some role in the spread of neurons that fire in response to an object, but neuroscientists contend that "the

potentially more important component arises from the fact that [objects] of the same width can look different, and thus give rise to different neuronal responses, when viewed from different distances or vantage points" (Ma et al., 2006, 1432). Depending on the acuity of the sensory modality involved, a larger or smaller set of neurons with a wider or narrower range of preferred stimuli will fire in response to an object. In a typical visual trial, neural activations for seeing the 70mm bar peak sharply, with high activation only for neurons whose preferred stimulus is very close to that mark; in a haptic trial, neural activations for touching the 70mm bar have only a softer rise near that point, with more diffuse activation. When the two modalities are used together, however, the combined neural signal is sharpened beyond that of vision alone, simply by summing the activations of each modality for that object: the small boost supplied by the touch signal at its soft peak gives the resulting combined curve higher gain and correspondingly lower variance (Ma et al., 2006, Figure 2).

For sensory integration, the subject must use these senses simultaneously on a single object: enhanced gains are not available to the subject who alternates periods of hands-off seeing and closed-eye touching. It is nontrivial to align the two sensory signals from an object in order to sum the resultant neural activity, but one vital aid to alignment is that the subject is an agent whose motor commands draw guidance from both sensory channels together in real time: a single motor map overlaps the various sensory maps (Stein, Stanford, & Rowland, 2014). As one explores the object itself on the basis of that sharply combined guidance, one receives error signals that are informative to both contributing modalities.

There is only a rough analogy between the way in which senses are united in a subject and the way agents are united in a group of the type that can gain common knowledge. Individual sensory channels never function as independent agents in their own right, and a unified human subject continues to exist while switching off individual sensory channels, for example by closing his eyes. A group such as a conversational dyad involves participants who are independent agents with their own private knowledge, and the dyad ceases to accumulate knowledge as a subject if either agent drops out, say by walking away out of earshot. But the productivity of sensory integration can inspire a search for ways in which human groups might be similarly productive, with the group's knowledge on a topic generally enhanced by the activity of all members. We can also make use of the idea that unification can happen in virtue of some shared engagement with the environment, as opposed to the formation of increasingly complex internal representations of representations (going back to the other side of the analogy, it is not as though each sensory modality initially receives stimulation from the world, and then privately constructs a series of increasingly diffuse representations of the neural activation in other modalities). Lastly, we can look for a relevantly similar connection to action as a unifier.

4. Common knowledge in the conversational dyad

To explain the unifying architecture supporting common knowledge, it will help to begin with the special case of the conversational dyad. The dyadic case is basic: we start to learn its rules in the first months of life, in the one-on-one 'proto-conversations' between infant and caregiver taking turns vocalizing and responding (Bateson, 1975; Yoo, Bowman, & Oller, 2018). By adulthood, our knowledge of these rules is well entrenched, with typical adults spending several hours every day immersed in conversation (Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007). Adult conversational groups still most often involve just two people: in social settings where larger conversational cliques are free to form, dyads predominate and cliques of more than four occur only rarely (Dunbar, Duncan, & Nettle, 1995). Even when conversations do involve multiple participants, they run largely on "serially dyadic participation" in which speakers focus primarily on one addressee at a time in a series of shifting pairs (Stivers, 2021).

As individuals, we can enter and exit dyadic configurations freely, but the basic structure of the configuration is fixed. There are two well-defined roles, for which I use the conventional terms *speaker* and *addressee*, while emphasizing that the addressee role also involves active vocal and gestural signaling, as opposed to merely passive recipiency. Conversation researchers point to this communicative activity as something that sets addressees apart from listeners who just happen to overhear a conversation (Bavelas, 2022, 3). Individual participants switch frequently and swiftly between the roles of speaker and addressee, and do so in a well-coordinated manner, with only rare collisions or confusions about who is occupying which role (Levinson, 2016). Both roles are vital for the formation of dyadic knowledge. Just as both senses must operate simultaneously on an object to achieve bimodal sensory integration, both speaker and addressee must act overtly on a propositional content in a sequence for the dyad to pass judgment on it, with the speaker generating the content, and the addressee evaluating it, paradigmatically at a moment of eye contact.

In the role of speaker, one has access to the full vocabulary of the language, as well as meaningful co-speech gestures: speakers alone can generate the topical content of the conversation, bringing it into joint attention (the 'main channel' of communication). Addressees, by contrast, deploy only a limited range of 'backchannel' signals, all of which have interaction-guiding or evaluative functions. These signals encompass verbal interjections such as *yeah, oh, mhm, no, exactly, okay, mm, uh, huh?*, and gestures such as nodding yes, headshaking no, expressions of puzzlement, smiling, and frowning. For reasons to be examined shortly, these backchannel signals are not themselves a focus of joint attention for the dyad, but they do register in the individual awareness of the participants. Because signals like nodding and *mm* are often omitted in transcription, and also because they are generally produced unreflectively and do not register in joint

attention, it is only with the help of with recent audiovisual corpora that we have become explicitly aware of the sheer frequency of these signals, occurring every few seconds in free-ranging conversation (e.g. Reece et al., 2023). This backchannel activity seems to be exclusively a feature of human communication: while many other animals signal each other, it seems none of them issue news receipt tokens like *oh*, continuers like *mhm*, or repair initiators like *huh?* (Dingemanse & Enfield, 2024; Dingemanse et al., 2015; Enfield, 2017).

Sequences are always started by speakers and ended by addressees. After greetings or attention-getters which launch and signal the formation of the dyad as such, speakers can either *tell*, *ask*, or *command.* In acts of telling, the speaker generates the topical content, bringing it into joint attention; the addressee can signal acceptance or rejection (*yeah, oh, okay, no,* nodding, headshaking), or more neutral comprehension (*mm, mhm*). In acts of asking, the speaker directs the attention of the dyad onto a question, obliging the addressee to switch into the role of speaker and answer with a telling. When speakers command, addressees may either comply in action, often prefacing that action with a verbal or gestural assent (*okay,* nod), or resist (headshake, *no*). For any of these moves, the addressee has a further option of requesting repair (*huh? pardon?*).

In some models of common ground formation (e.g. Stalnaker, 2014), the speaker's contribution enters common ground unless the addressee protests; such silence-is-consent models make it somewhat mysterious why addressees send out such a flood of positive comprehension and acceptance signals. By contrast, the integrative model calls for positive action on the part of the addressee. In ordinary cooperative conversation, it is only when the addressee responds to the speaker's statement that *p* with an acceptance signal (paradigmatically *oh, okay*, or *yeah*) that the dyad has thereby judged *p* to be the case, ideally making *p* commonly known (but sometimes just commonly accepted or believed). If the addressee is unresponsive, or if the addressee issues a more neutral signal like *mm* with falling intonation, then the dyad does not judge *p* to be the case, and the speaker's effort at a transaction has failed (apparent exceptions and special cases will be discussed below). The *mm* signal does however ratify that the dyad has jointly attended to this content: the fact that the speaker has said that *p* is now "out in the open" between speaker and addressee, even if the fact that *p* itself is not. If the addressee issues a repair signal like *huh?* or *pardon?,* then it is not even out in the open that the speaker has said that *p.* Moving down the evaluative spectrum to its negative pole, the addressee may reject the speaker's contribution with a headshake or *no,* but the addressee doesn't thereby make it the case that the dyad accepts not-*p*. As an addressee, one can either accept, request repair on, or rebuff the speaker's proffered content, but to establish its opposite one would need to switch into the role of speaker and state one's case. Because content always comes from one side and evaluation from the other, neither individual ever has lone control over what the dyad judges.

The suggestion that common ground emerges from activity on both sides is not new, but existing models that take this type of approach (such as Herbert Clark's 1996 presentation-acceptance theory) largely ignore the properly epistemic dimension of this interaction, sometimes going so far as to insist that the epistemic quality of interactively gained alignment is irrelevant, as long as speaker and addressee end up on the same page, or even just end up both believing that they are on the same page (see e.g. Clark & Krych, 2004, 63). Janet Bavelas compares interactive calibration in conversation to the synchronization of wristwatches between plotters who could well set them at any arbitrary time, as long as they make the same choice. "Similarly," she continues, "the meaning that any pair of interlocutors arrives at through calibration is neither correct nor incorrect; it simply works for them at this moment in their dialogue" (Bavelas, 2022, 151). Models of this sort face some challenges in explaining systematic differences between addressee signals of comprehension and acceptance. Attention to those differences enables us to see how dyadic configurations specifically tend to drive the formation of joint knowledge, a higher-value state than mere alignment.

In my model, backchannel overtly evaluates the speaker's contribution without itself placing new information into joint attention for the dyad. Because an addressee's *mhm* or *yeah* is not itself in joint attention, it doesn't itself call out for a further receipt signal or evaluation from the other side, stopping the threat of a regress of mutual evaluations. The dyad's attention is always on the main channel content contributed by speakers, although evaluative feedback from addressees works through individual awareness to determine how that content is handled. Speakers are individually aware of their conversational partner's overtly signalled state of confusion, comprehension, or acceptance, in a way that automatically guides their real-time behavior as speakers, even as the shared focus of the conversation remains squarely on the speakers' topical content and not on the addressees' struggle or satisfaction. If you send a repair signal, I need to repeat; if you comprehend but don't accept, I need to work harder to persuade you. Of course, either party is at liberty to use the main channel of communication to direct joint attention onto an addressee's predicament or even onto the particular backchannel signals just sent; the claim is only that backchannel signals do not naturally command joint attention on their own.[2]

[2] A similar division of labour between joint and individual attention is proposed in recent linguistic anthropology. Against earlier work suggesting that pragmatic features of language lie beyond the awareness of conversational participants, Charles Zuckerman and Nick Enfield (2023) argue that these features do fall within individual awareness, but not within joint awareness: they are not naturally 'thematized' just by being used, although participants can choose to thematize them (bringing them into joint awareness), when given suitable resources and motivation through interaction.

This understanding of backchannel builds on earlier theories. The first conversation theorist to use the term was Victor Yngve, who introduced it as follows: "both the person who has the turn and his partner are simultaneously engaged in both speaking and listening. This is because of the existence of what I call the back channel, over which the person who has the turn receives short messages such as "yes" and "uh-huh" without relinquishing the turn" (Yngve, 1970, 568). In Yngve's view, backchannel consisted of verbal response particles signalling the addressee's interest or attention during an ongoing speaker turn. Speakers expect these signals, and are distressed by their absence: just stop issuing them on a long-distance phone call as another is speaking, Yngve says, and the speaker "will soon come to a grinding halt and say something like "Hello, are you still there?"" (Yngve, 1970, 568).

Yngve sees backchannel as addressee vocalization during an ongoing speaker turn. I take backchannel to encompass all evaluative signals the addressee sends the speaker both during and immediately after the speaker's contribution, including non-verbal signals such as nods, headshakes, most smiles, and expressions of puzzlement. While Yngve is certainly right that *uh huh* can be uttered without the speaker's relinquishing their turn, say, giving a storyteller the green light to continue holding the floor, later theorists recognized that backchannel signals also appear at moments of speaker turn completion, either as free-standing turns on the addressee's part, or prefacing the addressee's transition to speaker. Free-standing backchannel turns include for example uses of *oh* when a question has been asked and answered (Heritage, 1984); as sequence closers, these addressee receipt tokens do not call for a further response from the other side. In the transition to taking the floor as a speaker, individuals often preface their turns with what I classify as a backchannel signal evaluating the prior turn, still from the position of addressee. It has been estimated that roughly half of new turns in spoken English start with "something besides a constituent of a grammatical unit" (Norrick, 2009, 871). Neal Norrick's corpus-based list of these most common "non-grammatical constituents" extends beyond backchannel (including for example the speaker hesitation sounds *uh* and *um*) but the two most frequent starts are *yeah* and *oh*, both of which have the main function of evaluating the prior turn, and when they are used in that way I classify them as backchannel from the addressee side, with individuals entering their own turns as speaker only after this backward-looking evaluation is out of the way.[3] Backchannel signals are typically sent at moments of syntactic completion, with continuers such as *mhm* appearing more at the end of clauses such

[3] The third most common turn-starting "pragmatic marker" on Norrick's list is *and*; it may be surprising to see this conjunction put forward as something that is "not a grammatical constituent" of the utterance. However, in initial position, the conjunction can be read as a backward-looking endorsement of the turn just delivered by the other party. With that backchannel acceptance out of the way, and the initial conjunction is indeed not a constituent of the new turn the incipient speaker is producing; rather, it sets this turn as up as building on what came before.

as the antecedents of conditionals, and change-of-state signals such as *oh* and *okay* appearing more at the end of sentences and turns, especially where falling intonation indicates that the speaker is wrapping things up (Guthrie, 1997).

Backchannel stands as an exception to the general rule against producing and consuming speech at the same time: ordinarily, performance as a speaker collapses if one must simultaneously listen to other speech, and listening comprehension collapses if one must at the same time talk (Jaffe, 1987).[4] By contrast, speaker performance is enhanced by the presence of an addressee freely producing backchannel while one is talking (Bavelas, Coates, & Johnson, 2000; Krauss & Weinheimer, 1966), and an addressee producing backchannel will understand a speaker better than a paired eavesdropper just overhearing the conversation (Schober & Clark, 1989).

Several features of backchannel communication help to explain how it can be processed alongside the main channel. Backchannel signals largely occupy a narrow and distinctive bandwidth: these signals are sharply limited in what they can express, and somewhat restricted in how they express it, with signatures of these limits appearing even at the phonetic level. To take English as an example, while speakers use the full lexicon, spanning roughly 40 phonemes, addressees heavily favor a smaller range of signs, many of them "non-lexical conversational sounds" such as *mm, uh-huh* and *mhmm* drawing from a smaller set of just 10 phonemes (Ward, 2006).[5] Some backchannel sounds are distinctive—for example, the *tsk* of disapproval is a click sound not otherwise found in English (Ameka, 1992)—but even common sounds can be deployed in uncommon ways when they appear in backchannel. In his treatment of the "response token" *mm,* Rod Gardner writes:

> *M*m as a sound, as the phone [*m*], has some characteristics that make it distinctive in certain respects in English (and in many other languages). As a bilabial nasal continuant, it is the only sound in English which has the mouth closed from onset to termination. Further, there is no lip, tongue, or jaw movement during the production of the sound, i.e. there is essentially no movement of the mouth and jaws associated with its production. The low prominence of the auditory production is thus complemented by a visual message that the producer of these tokens is presenting minimal vocal

[4] Simultaneous interpretation of the type performed at the United Nations is extremely difficult, imperfectly executed, and not exactly simultaneous. To the extent that it succeeds, it seems to be enabled by bursts of productive work during the speaker's pauses, aided by the fact that some of what the speaker is saying is predictable or ignorable, and by the fact that the interpreter's incoming and outgoing streams of speech should have the same semantic content (Christoffels & De Groot, 2005).

[5] Ward's classification of these signals as "non-lexical" is controversial, with some theorists arguing that it "detracts from the systematic, conventionalized nature of these items" (Dingemanse 2023, 478). The phonetic distinctiveness of backchannel is not simply an English phenomenon; researchers note similar patterns across many languages (e.g. Ameka 1992, Dingemanse 2004).

> activity and labial closure. The sound [*m*] is the only one in English that carries all these minimal characteristics. (Gardner, 2001, 66-7)

The distinctive minimality of backchannel phonetics is apparent in common repair signals as well. Addressees can use a word like *huh?* to prompt speakers to repeat themselves. Mark Dingemanse and colleagues have found striking cross-linguistic commonalities in this signal: surveying a wide range of languages, they find that in all of them this signal appears as a single syllable with no final consonant, an unrounded mid-low vowel, and whatever intonation the language associates with questioning (rising in most languages, including English, but falling in languages such as Icelandic) (Dingemanse, Torreira, & Enfield, 2013). They explain these features as the products of convergent cultural evolution in response to the universal need for a quick signal of comprehension failure; the pattern they find minimizes articulatory effort while still constituting a distinctively questioning sound in the relevant language. The same principle extends even to Argentinian Sign Language, they observe, where eyebrow raising is generally used to signal questioning, and *huh?* is indicated with just a minimal eyebrow raise. Dingemanse and colleagues suggest that convergent cultural evolution may also explain broad cross-linguistic similarities in other distinctive backchannel signals such as the continuers *mhm* and the epistemic change of state tokens *oh* and *ah* (Dingemanse, 2023; Dingemanse et al., 2013). For example, the Mandarin *ou* is similar to *oh* not only in its function of signalling knowledge gain, but also in its sound and declarative intonation (Wu & Heritage, 2017). There is some diversity here across languages, for example in Finnish, where *aijaa, aha(a)* are both heavily used to do the work of *oh* (Koivisto, 2016), but we have no attested language without some simple set of signals that can be used from the addressee position to request repair, signal ongoing attention, or mark acceptance or knowledge gain, all in response to what the speaker has said.

Phonetically minimal backchannel words like *mhmm, yeah* and *oh* are not only the most common interjections, but among the most common words of any class in conversational language use (Dingemanse, 2024, 258). However, backchannel goes beyond these 'special-purpose' tokens to include words from the larger lexicon: a recent corpus of American English conversation also includes ordinary words such as *exactly, nice, right, sure* on its list of most frequent backchannel items (Reece et al., 2023, Supplementary materials p.12). Traditional theories characterize these as 'secondary' interjections, as opposed to the primary interjections like *mhmm* that are not otherwise used in the language. Both types of interjection are distinguished by the way they appear in discourse as uninflected, non-clausal, non-elliptical utterances (Ameka, 1992, 105). Prototypical interjections are single words, although backchannel utterances can string words together in short phrases, for example combining a news receipt token with an evaluation (*oh wow, oh dear*) or compiling more emphatic multi-word

acceptances (*yeah right exactly*). These concatenations still lack propositional structure, however, ensuring backchannel utterances remain informationally lean enough to be produced and understood simultaneously with main channel speech.

Across languages, backchannel is heavily used, although comparative inferences are difficult because of differences in the ways researchers classify backchannel and collect corpus data. A review looking at 1.3 million turns of dyadic conversation in transcribed corpora from 18 different languages reports 'single unit interjections' as occurring once every 12 seconds (Dingemanse, 2024). For statistical purposes, because the corpora were different sizes, this count was restricted to the top ten most frequent verbal tokens in each language, and further restricted to single token turns (so including a standalone turn of *oh*, for example, while excluding *oh good*). If we drop those two restrictions, we see higher counts: the 8-million word CANDOR corpus of Zoom conversations finds about a thousand addressee backchannel words per hour, one every 4 seconds (Reece et al., 2023, p.7). If we include not only vocalization but head movement, rates are higher still, with addressees in the CANDOR videos nodding yes or shaking no roughly once every three seconds. Corpus participants privately rated the quality of their partners as conversationalists after these video calls; addressees whose ratings put them in the top quartile of conversationalists were more expressive in their backchannel, showing both higher average rates of both nodding yes (25% of turns) and shaking no (21% of turns) than those rated in the bottom quartile, who nodded at 21% and shook their heads at 18% of speaker turns (Reese, Supplementary materials, figure S.2, p.27).

Speakers not only enjoy expressive and differentiated backchannel, but seem to seek it out. One measure of backchannel quality is the ratio between what Janet Bavelas terms 'generic' backchannel (*mhm, yeah*) versus 'specific' backchannel (*oh, wow*) signalling interest in the particular content just delivered. During extended dramatic story-telling turns, addressees produce these tokens every 3-4 seconds, at a ratio of about two generic tokens to one specific. In an experimental condition, speakers had to tell a 'close call' story of narrowly averted danger to an addressee who was distracted by the task of pressing a hidden button every time the speaker used a word beginning with the letter *t*. Distracted addressees dropped to producing less backchannel overall, and virtually all of it generic. Speakers who faced these distracted addressees told worse stories, badly paced and rambling, with choppy endings and fumbling, repetitive attempted justifications of the story as an illustration of danger, as if attempting to elicit an *oh wow* (Bavelas et al., 2000). Closer analysis of the interplay between ordinary backchannel and speaker behavior shows that when storytellers receive a generic continuer signal like *mhm* or *yeah*, they tend to go on to the next chronological event, and when they receive a specific reaction like *oh*, they tend to elaborate on what they have just said, a pattern confirmed by experimental work in which

these tokens are artificially swapped (Tolins & Fox Tree, 2014). The addressee's *oh* gives the speaker live feedback on what the addressee finds surprising, a signal whose epistemic value will be explained shortly.

One last psychological feature of backchannel warrants attention before we turn to questions of epistemic value. In face-to-face conversation, there is an intimate relationship between backchannel and eye contact. As a general rule, addressees keep their gaze focused on the speaker, while speakers mainly avert their gaze as they speak, selectively looking at the addressee as they end their turns, but also at moments of soliciting feedback in an ongoing turn, often at points of syntactic completion (Kendon, 1967). Each moment of eye contact is a "gaze window" that the speaker opens by glancing at the addressee, and closes by looking away after the backchannel signal is received, if the speaker is taking an extended turn (Bavelas, Coates, & Johnson, 2002). Ending one's turn as speaker with direct gaze at the listener is a way of cueing her to start speaking (Ho, Foulsham, & Kingstone, 2015), but the brief overlap of mutual gaze between the parties also allows the backchannel evaluation of the last turn.

Eye contact has a power rooted in a deep evolutionary history for us: it is "an unambiguous stimulus with tremendous evolutionary significance", triggering deeply embedded neural responses across predator and prey species alike (Shepherd, 2010, 1). The critical moment of simultaneously seeing and being seen naturally has mutual significance for conspecifics as well, in a way that circumvents the need for recursive representation, given the parallel visual equipment of the two agents, their shared innate instincts, shared background of social learning, and similar activation in eye contact. When two primates look directly at each other, each simultaneously receives the same type of stimulus, triggering the same powerful innate and learned mechanisms for both animals at once: they are synchronized in their visual awareness of each other. Among primates, humans have heightened sensitivity to eye direction, with our dark irises standing out in sharp contrast against the uniquely human exposed white sclera (Kobayashi & Kohshima, 2001). Humans are exquisitely sensitive to direct eye contact, even as newborns (Guellaï, Hausberger, Chopin, & Streri, 2020), and we have moments of eye contact roughly every two seconds in face-to-face dyadic conversation, with roughly two-thirds of those moments happening at the end of turns (Wohltjen & Wheatley, 2021). These moments of eye contact form safe corridors for backchannel signals.

For a conversational dyad to judge that *p*, they must achieve shared attention to *p*, initiated by the speaker's sincere signal that *p*, to which the addressee must then sincerely and successfully signal acceptance.[6] Insincerity on either side creates the mere illusion of dyadic judgment. Individual judgment can come apart from dyadic judgment in various ways, most obviously in main channel lying, where it will not be the case that the dyad together judges something the speaker privately judges to be false. Backchannel signalling can also be insincere; I can pretend to accept something just in order to placate you. In a subtler case, your telling me that *p* might cause me to learn that *p*, but if for some reason I do not wish to acknowledge the truth of *p* publicly, I can refrain from nodding, and instead issue a skeptical-sounding *mm* signal with falling intonation. This outward show of resistance serves to block the fact that *p* from being "out in the open" between us. Thanks to your telling me, the fact that *p* is something we now both know, and indeed you might even privately know that I now know it, as you read some involuntary flicker of unease on my face, but we do not know it as a dyad; it is not common knowledge between us.

The signalling loop between speaker and addressee must be successfully completed for the dyad to reach judgment. If the speaker does not happen to see or hear the addressee's confirmation, then the addressee may be under the illusion that the dyad has judged that *p* when this is not the case. However, several mechanisms work to ensure that bidirectional signalling is normally successful. In face-to-face conversation, unsuccessful addressee signals are rare, because these signals draw from a simple and familiar register and happen mainly at moments of eye contact. Speakers expect appropriate backchannel from addressees and will repeat themselves if they do not hear it. When the speaker builds on a point that has just won addressee assent, rather than repeating, this confirms back to the addressee that her signal was received.[7] Illusions of dyadic judgment will typically be short-lived: unsuccessful signalling in either direction triggers repetition or repair sequences, some of which involve reversal of the speaker-addressee roles.

Across languages, addressee-initiated repair sequences fall into three types. An 'open request' signal such as *pardon?* or the minimal *huh?* is sent when the addressee misses the entire speaker signal. If only part of the

[6] Measurements of that shared attention show an interesting relationship to eye contact. Under conditions of constant illumination, pupil dilation correlates with attention, even when the stimulus is not visual but, say, musical or linguistic. The level of dyadic synchrony in pupillary response is therefore taken as a measure of the degree of shared attention in the dyad. In natural conversation, eye contact does not trigger shared attention; rather, high shared attention seems to trigger eye contact, and that shared attention stays high during the brief moment of addressee confirmation. Shared attention then drops as the next speaker looks away and introduces fresh content (Wohltjen & Wheatley, 2021).

[7] A study of get-acquainted conversations found 97% of turns introducing new information gained this kind of 3-step overt calibration between speaker and addressee (Bavelas, Gerwing, & Healing, 2017).

signal is lost, the addressee sends either a 'restricted request' (*It starts when?*) or if part of the signal is uncertain, a 'restricted offer' (*at 9 o'clock?*). Experimental work has shown that rather than sticking to a minimal-effort *huh?* for all repairs, addressees systematically follow the principle of least collaborative effort, selecting whichever repair signal (open request/ restricted request/ restricted offer) will minimize the total work of the dyad (Dingemanse et al., 2015). Reinforcement learning offers a straightforward explanation of this pattern of behavior: while issuing a minimal *huh?* is always the addressee's lowest-effort single action in the moment, a fixed policy of open requests would often waste time for the dyad (including the addressee's own time) by prompting the speaker to repeat the whole turn. If the addressee is missing only one word, issuing a restricted request means shouldering more of the immediate burden of the repair, but delivers a swifter overall result. Over time, reinforcement learning (RL) supports exactly this kind of switch from policies yielding small immediate gain to policies yielding larger longer-term gain. Naturally developing automatic habits that reflect the expected longer-term average reward save us from having to think strategically about our choice of repair strategy (for a review of the relevant features of RL, see Haas, 2022). These learned habits can also include switches between the roles of speaker and addressee. In extending a restricted offer, the addressee bypasses backchannel delivery to take a turn as speaker, allowing the original speaker to assume the role of addressee in evaluating the restricted offer. The dyad's work is done on the proposition when bidirectional signalling has occurred on it, regardless of whether it is the original addressee or speaker who issues the final evaluation.

A dyad judges that *p* by closing a loop of sincere signalling on it. Dyadic judgment guides dyadic action, governing both what the dyad goes on to do in conversation, and how the dyad acts in the world. The dyad can act only through the bodies of its individual participants, but in being guided by dyadic as opposed to merely individual judgment, their actions are coordinated. In judging that *p,* the dyad's attention is on the main channel content *p,* as opposed to the fact that the dyad is now judging that *p,* but the fact that this judgment is dyadic in character falls within the individual awareness of each of the participants, thanks to their well-rehearsed individual knowledge of the structure of conversation, and the bilateral overt signals activating this structure. When a dyad judges that *p*, both parties can act on *p* in a synchronized way without needing to engage in recursive mindreading, given their possession of parallel cognitive equipment simultaneously activated by the same two-step (main channel, backchannel) sequence of stimuli. Like the prison in Section 2, dyadic judgment is a structure that unlocks only when two different buttons are pushed at the same time. The claim of simultaneous activation may seem inconsistent with the sequential character of the paired stimuli; however, it is essential that the main channel activation of content remains in joint attention through the moment of its backchannel evaluation. A backchannel signal like *mm* has no meaning on its own, but only as coupled to a main channel content.

Even if its natural function is to support dyadic knowledge, this cognitive equipment can be also activated by problematic stimuli. Insincerity on either side creates the illusion of dyadic judgment, rather than the real thing, and this illusion may be enough to motivate the duped party to act contrary to his interest. However, larger forces of reinforcement learning ensure that signalling is typically not only successful but sincere. In contexts where the relevant interests of the two parties are aligned, there is no incentive for insincerity. In contexts where interests diverge, a hostile party can produce the illusion of dyadic judgment that *p* through insincere signalling, but to the extent that it is to the deceived party's disadvantage to act on *p*, we can expect conversational participants to get better, over time, at differentiating such contexts. Over time, the costs of being fooled by liars sharpen one's capacity to resist them, for example by increasing sensitivity to situations where interests diverge, making individuals more distrustful in those contexts.[8] Lying is also generally more taxing than sincere signalling: to generate the illusion of knowledge for the other side, liars must tie up attention to secure the internal consistency of that constructed illusion and keep it quarantined from their own grasp of reality (Westra & Nagel, 2021). Meanwhile, those who are often insincere risk being shunned as conversational partners, as the costs of trusting them mount up and form a topic of discussion in the larger community; considerations of social reputation and potential loss of signal power also protect the sincerity of signalling.

Sincerity on both sides is not enough for dyadic knowledge. Speaker and addressee can whole-heartedly agree on something that is false, or only coincidentally true.[9] But although the architecture of dyadic judgment can work to synchronize agents to act on a falsehood or coincidental truth, it is a mistake to conclude that this architecture runs in a way that is indifferent to the line between factive and nonfactive mental states. Over time, if it is systematically more rewarding for parties to synchronize on the truth, they should learn to signal, both as speakers and as addressees, in ways that systematically close the loop on signaling that *p* only if *p* is the case. It is easiest to see how this kind of learning will work in conversations whose point is to plan immediate joint action: to the extent that plans tend to be frustrated by false dyadic judgments and furthered by true ones, our backtracking calculations of value at the later moment of negative or positive reward will favour our

[8] This does not mean that we get better at spotting liars in artificial laboratory environments, where confederates are instructed to lie about arbitrary subject matter, as opposed to being motivated to lie in ways that serve their (presumably more predictable) real world interests.

[9] In addition to sincere dyadic judgment falling short of knowledge, they can also in all sincerity create the illusion of dyadic judgment, for example where the addressee signals assent on the basis of a mishearing of what the speaker has said. Because such illusions do not reliably coordinate individuals into joint action, the same action-dependent factors that favour synchronization on the truth will also work towards stamping out illusory dyadic judgment, over time.

signalling in ways that must latch the dyad onto the truth, assuming that there are learnable regularities we can exploit here (cf. Berke, Walter-Terrill, Jara-Ettinger, & Scholl, 2022). Some relevant regularities are mapped out in the sociological literature on epistemic territory: we routinely expect others to be knowledgeable about their "thoughts, experiences, hopes, and expectations", and domains such as their "relatives, friends, pets, jobs, and hobbies" (Heritage, 2012, 6). So for example, in 'get acquainted' conversations, addressees rapidly accept a newly-met stranger's claims to play the French horn, or to work for a building supply company (Bavelas et al., 2017; Reece et al., 2023). The epistemic legitimacy of the addressees' assent can be grounded in their recognition of these claims as falling within the kinds of topical zones in which speakers are generally knowledgeable. Over time, addressees gain a better sense of these zones by interacting with multiple speakers in multiple contexts, backtracking to update their maps of epistemic territory when they later encounter trouble or satisfaction.

In order to signal dyadic acceptance, the addressee needs no prior independent evidence that some new stranger who says she plays the French horn does in fact do so, if this claim falls within a learned type of trustworthy claims. For claims in more contested or controversial areas, speakers will need to work harder to make their knowledge recognizable, for example by offering arguments starting from premises that the addressee will take on trust (Mercier, 2016).[10] However, even within a speaker's usual zones of expertise, there may be points at which the addressee knows she has a superior epistemic position—say, I saw your pet struck by a car just prior to our conversation—enabling that addressee to override the usual topical deference and stop the dyad from accepting a falsehood.[11] Because claims must pass through both parties' epistemic filters for dyadic acceptance, the dyad will have higher epistemic standards than either individual party, assuming some baseline of epistemic competence on both sides.

To a first approximation, sincere dyadic acceptance of a speaker's claim that *p* can be modeled as the addressee's recognition of the speaker as knowing that *p*, where the addressee's response can draw on any generalizable features of the larger situation, including the specific propositional content and the addressee's perceived relation to that content. However, if the integrative model is correct, the speaker does not necessarily need to have individual knowledge that *p* in order to produce dyadic knowledge that *p* by placing *p* in joint

[10] The RL framework can also explain our gradual mastery of valid argument forms: to the extent that we are rewarded, as dyads, for converging on the truth, we should expect speakers and addressees to learn to produce and accept the same truth-preserving patterns of operations on propositional contents.
[11] Of course, it is also possible for the addressee to have some misconception that bars the dyad's acceptance of a truth known to the speaker, a result which would typically drive the speaker to make more of a case for it.

attention. If it is rewarding for dyad to act on truths, the speaker just needs to signal in ways that will systematically yield acceptance that *p* by the addressee only if *p* is true. Speaking to an epistemically better-positioned addressee, I might venture a hedged remark, hoping for the kind of *yes exactly* that could upgrade my tentative confidence to dyadic knowledge, either as a function of your pre-existing knowledge, or as a result of your independently high confidence that needed just a small boost from me to ground outright judgment for both of us.[12] Over time, if our practices of combining signals are appropriately shaped by feedback from the world, dyadic knowledge should be attainable on some points where neither party clears the threshold for knowledge on their own.

Because parties in instrumental conversations are planning to act together, they have an obvious motivation to upgrade their individual judgment to dyadic judgment, and a good prospect of getting feedback from the world on the quality of that judgment. Having agreed to meet for lunch an hour from now at a certain restaurant, we will find out soon enough whether we were right to judge together that it was open today, and the sting of disappointment on discovering that it is closed will naturally shape our future practices of judgment on such points. However, most of our conversations are not transactional: action planning is estimated to take up only about ten percent of adult conversational time (Dunbar, Marriott, & Duncan, 1997). Across cultures, the bulk of adult conversations have no clear consequences for bodily action (Dunbar et al., 1997; Emler, 1994; Haviland, 1977). These "inconsequential" conversations might seem at first to present a problem for RL-based theories of mutual signaling. However, closer examination shows that idle gossip also hones the epistemic quality of our signalling, with intrinsic as opposed to instrumental reward playing the key role this time.

Coordinated bodily actions like meeting for lunch depend on the specific judgments we have made as a dyad, but these are not the only actions we perform. After greetings, speech actions within a dyad depend directly on prior dyadic judgments. I tell you something, and if you accept it, we can both build on it; if you resist, I must change course, perhaps working harder to establish my point, perhaps shifting topic. Given the forward-looking character of action, both parties naturally form expectations about what the dyad will accept as the conversation is unfolding. In part because neither party fully controls what the dyad accepts, conversation poses an interesting challenge to our powers of prediction, and we are often surprised by the turns it takes.

[12] This approach is in line with Peter Van Elswyck's (2023) view, according to which speakers who issue an unhedged declarative that *p* can be understood as signalling that they know that *p,* and hedging signals a lesser epistemic state.

The reward we find in these surprises arises from a natural human characteristic closely connected to knowledge: curiosity.

Traditionally defined as an intrinsic desire for knowledge gain, natural curiosity can be modelled within the RL framework as an appetite for surprise. I have argued elsewhere that these two characterizations are equivalent (Nagel, 2024). Because our most basic learning mechanisms are driven by prediction error, there is an intimate relationship between surprise and learning. We learn nothing when everything happens exactly as we expected, but when our expectations are violated, we experience surprise and update our models of reality. Surprise functions as a marker of an educational experience, breaking through to conscious availability when the violation of our expectations is significant. This marker enables curious creatures like us to have a kind of meta-learning: as we learn about the world, we simultaneously learn what types of circumstance-action pairings yield significant updates to our models of reality. As the world becomes better known, we need to enact longer and more sophisticated action sequences to find surprise. Creatures driven strictly by extrinsic motivations such as thirst and hunger will also learn from surprising experiences, but creatures who are attracted to surprise as such will be motivated to act in ways that accelerate knowledge gain, for example by venturing into unexplored areas and engaging with challenging stimuli. The fact that we find intrinsic reward in surprise drives us to antagonize our own models of the world, testing them at remaining points of uncertainty. The internally adversarial interaction between prediction-error correction processes of learning and prediction-error-seeking appetite of curiosity accelerates our cognitive adaptation to reality, driving the formation of representational states whose stable existence depends on their truth, states of knowledge.

The fact that curiosity drives us to engage with hard-to-model aspects of our environment explains something of the attraction of conversation as an activity: other humans are typically the most complex objects in our environments,[13] and conversation allows us to explore that complexity. The fact that we experience reward from surprise makes conversation pleasurable for us, and the fact that we can learn what is surprising from backchannel feedback spares us from having to make taxing strategic calculations.

[13] Perhaps setting aside our smartphones, whose complexity and attraction for us itself stems only in part from their role in mediating various types of social conversation. Interactions with these devices involve such a complex mix of conversational and other rewards that I cannot aim to cover them in this article. For present purposes it is enough that our language use still starts with dyadic face-to-face conversation, and we continue to engage in this type of communication quite often, even in contexts where smartphones are also available.

Patterns in backchannel reveal an interesting role for curiosity in non-instrumental conversations. One puzzle in backchannel concerns the addressee's choice between the acceptance tokens *oh* and *yeah*. The standard theory of *oh* is that the addressee uses it to signal knowledge gain, or the retrieval to current awareness of previously acquired knowledge (Heritage, 1984). We do not usually say *oh* upon being told something already known;[14] the standard response to hearing something one already knows is *yeah*. But not all new items of knowledge merit an *oh*. Conversation transcripts show many instances of *yeah* in response to what must be new information: for example, near the beginning of a CANDOR corpus conversation between randomly matched American strangers, a man named Chris gets a *yeah* in response to his claim that he lives in Wichita, Kansas, and another *yeah* for his claim that he works for a construction supply company (Reece et al., 2023, 8). One relevant feature of these claims is that they are relatively unsurprising: the addressee presumably had no particular expectation about the hometown or occupation of this randomly assigned partner, so learning these facts about him occasions just a minimal update of her model of reality. What triggers an *oh* is not simply novelty, but something more like a plot twist, with a considerable violation of expectancy; consciously available surprise marks more significant knowledge gain.[15] In get-acquainted conversations, speakers may need to share a number of background details in order to build up some expectations on the part of an addressee, to get into a position where some new detail is now quite surprising against the expectations generated so far.

A similar dynamic appears in story-telling, whether among friends or strangers. In story-telling, a speaker takes an extended series of turns. In the set-up phase, the speaker can give a number of details that will be novel but unsurprising to the addressee ("I was on the bus this morning, and it was crowded"). The addressee's *yeah* or *uh-huh* marks comprehension and acceptance of these relatively unsurprising facts, cueing the speaker to move forward to the next event. However, story-telling has a well-recognized structure: the narrator must build towards some point of surprise, some twist or climax. The story cannot simply consist of the crowded bus following its usual route and getting the story-teller to work more or less on time; especially as the story gets longer, we expect it to culminate in some kind of *oh wow* moment, and will feel disappointed if it does not.

The idea that non-transactional conversations are curiosity-driven explains this pattern. If surprise serves as reward in such conversations, story-telling is a learned strategy in which addressees endure a number of

[14] Unless perhaps one wished to convey the impression of just then hearing it for the first time, an abuse of the signal parasitic on its natural meaning.

[15] This violation of expectancy need not be particularly dramatic: one might respond *oh* on learning which of two candidates got a certain job, where one's prior credence was evenly distributed between them. Curiosity is thought to peak for questions on which one is 50/50 because these mark the highest expected information gain (Kang et al., 2009).

relatively unrewarding steps to gain a larger reward (the reward for the speaker is another matter, to be handled shortly). This drive for surprise enhances the epistemic efficiency of conversation more powerfully than the weaker norm not to tell others what those others already know. Speakers could comply with that latter norm by telling others endless small but privately held facts from within their recognized epistemic territories, down to what the speaker had for breakfast and when the speaker last brushed his teeth. As naturally curious animals, we receive intrinsic satisfaction from more surprising updates that occasion larger changes to our models of the world (say, I just discovered that our outwardly mild-mannered colleague is plotting something nefarious). However, if what I am telling you runs contrary to expectations, I may need to lay some groundwork both to establish your expectations, and to situate myself as sufficiently well-positioned on the key point to get you to accept my surprising claim rather than rebuffing it as improbable. I can do this by securing dyadic uptake on the necessary backstory first (our colleague was on that crowded bus, and didn't see me, but I could overhear what he was saying). The availability of a scale of smaller and larger acceptance tokens (*uh-huh, yeah, oh, oh wow*) enables this kind of progression towards sharing an exceptionally educational moment.

With limited conversation time, it is rewarding for addressees to hear a speaker's most surprising stories; these tend to carry higher epistemic value. But we might wonder why conversationalists in search of surprise do not simply play games with each other, as opposed to talking about the world. As a matter of fact, we start out that way: proto-conversations are pure gaming, starting in random play on the part of the infant, and even at 12 months of age, fully half of infant vocalizations are game-based or performative, as opposed to referential in character (Ninio, 2018, ch.6). Over the next two years, infants shift to referential talk about the here-and-now, and then about the larger world, with their percentage of game-based wordplay dropping to under 10% by 32 months of age, on its way down to adult levels of rarity. Once the game has been mastered, the reservoir of surprise available from variable timing in *peek-a-boo* is small compared to the surprise available in the larger game of making sense of the world, where we have increasingly rich and detailed expectations about how things are.[16]

So far, the focus has been on the rewards available to the addressee, but if conversation is to be a stable strategy, we also need to explain the speaker's incentives in non-instrumental conversations. The simplest path

[16] Surprise can also be delivered in fictional stories and narrative jokes, but these require more taxing strategic planning on the part of the speaker, and while they can involve some measure of world-building, significant reversals of expectation will typically need to happen against the background of the larger body of shared real-world knowledge ("it's funny because it's true").

is to model speakers as curiosity-driven, too.[17] Curiosity of course motivates speakers to ask questions, especially on points where the addressee is seen as knowledgeable and the speaker is at 50/50, for maximum anticipated surprise. But curiosity also motivates telling: speakers will be motivated to selectively share their most educational experiences if they learn by doing so. There is evidence that running through my most striking experiences will already have private epistemic value for me, even before I secure your reactions: processes of reinforcement learning are enhanced by the rehearsal, in episodic memory, of events that violated expectations (Gershman & Daw, 2017). The repeat global broadcast of these events, even just in inner speech or visualization, helps complete the work of updating my larger picture of the world. If I share the story with you in outer speech, I can call in your epistemic resources as well as my own; perhaps you can contribute further news of our unexpectedly villainous colleague, or supply some further detail that will enable us both to understand what is really going on.[18] The fact that I was surprised by what I heard on the bus is a cue for me that you will also find surprise there, given some background awareness of our similarity; it is also a sign of a weak spot in my model of the world, and perhaps in yours as well. Responses are attractively unpredictable at the weak spots, where we share partial knowledge. By pushing my surprising story into the realm of dyadic knowledge, I summon the power of the dyad to make sense of this raw territory together; a better-informed dyad will be better able to deliver results I myself will find surprising, working to repair this weakness with our combined epistemic strength.

5. Common knowledge as a default condition

There is a large gap between this partial account of dyadic face-to-face conversation and a general account of common knowledge covering more than two participants and other modalities of communication. In particular, extensive work is required to cover situations where backchannel has a different form, or where it seems to be absent. Some cases of common knowledge formation seem to feature roughly similar bidirectional signalling; notably, in joint visual attention, the initiator selects something for attention by pointing (or by gazing pointedly), and then the receiver evaluates the target with a facial response indicating attention (with or without surprise, pleasure, or displeasure), rapidly coupled with a moment of eye contact. Phone calls and texting arguably feature adaptations of the schemas learned in face-to-face conversations (Bavelas, 2022). Even

[17] This is not to deny that speakers also derive some reward from pleasing their addressees, making themselves more socially desirable in the process, and setting up expectations for reciprocity. Conversation has an affiliative dimension, involving the alignment of motivational as well as epistemic states. However, setting imperatives aside, conversational turns with clausal content are essentially informative (Heritage 2012), and not necessarily affiliative. To the extent that we are affiliating, we are doing so by means of joint participation in an informative exchange (as opposed to doing so by means of some other activity, such as holding hands or stroking each other's hair). Thanks to Johannes Mahr for helpful conversation on this point.

[18] Thanks to Sara Aronowitz for helpful conversation on this point.

in face-to-face conversation, we find adaptations of the simple system described in section four—a fuller account would acknowledge that we sometimes engage in processes of what Herbert Clark calls 'downward inference' to skip stages in backchannel by making moves that presuppose dyadic acceptance (Clark, 1996). The most challenging cases for my approach are those in which there seems to be no relevant interaction between the parties at all: for example, an announcement made over a loudspeaker seems to generate common knowledge among the people in a waiting room even if they neither look at each other nor vocalize any response.

The larger project of a more complete account of common knowledge is a project for another occasion, but that last case merits attention on this occasion, because it is one of the examples Williamson uses in arguing that we have a default stance of taking others in our environment to know salient truths just as we do, a "shared world" default. If this default stance is itself a salient truth, Williamson observes, then "the knowledge default is implicitly a *common* knowledge default" (Williamson, 2024, 46). He notes that on a recursive understanding of common knowledge, if we start with the schema "If P, everyone knows that P", and we substitute "Everyone knows that P" for P, we can generate "arbitrarily many iterations of 'everyone knows that'" (ibid). However, having just reminded us of his own arguments against recursive common knowledge, Williamson is clear that this schema is not to be taken at face value. Highlighting the gap between the attitude of knowing and the weaker attitude of taking a default stance, Williamson continues:

> Of course, such a default does not mean that there really is [recursively characterized] common knowledge. It just means that when nothing inhibits the default, everyone acts as if everything were common knowledge. But that may suffice for coordination to be achieved. It may even be achieved, just as it often seems, with no iteration of epistemic operators, indeed with no epistemic operators at all: the phenomenology is just that of a world open to view. Since the coordination is the predictable result of deeply rooted forms of human thinking, it may even be safe enough for those involved to know in advance that they will coordinate. (Williamson, 2024, 46)

The idea that we coordinate as the predictable result of deeply rooted forms of human thinking is entirely compatible with the RL-based model of common knowledge presented here: a history of both instrumental and curiosity-driven conversation supplies us with strong instincts about what others will and will not know, producing automatic expectations about their behavior without the need for reflective attention to their mental states as such. When in the company of like-minded others, we may knowledgeably expect coordination. As in communication, what is open to view is the main-channel material, whose acceptance governs our actions without this acceptance itself needing to be a focus of further attention: together, we typically keep our eyes on the world, and not on any epistemic operators.

Williamson is clear that his knowledge default should not be read as a suggestion that most agents are seen as knowing most truths, even restricting this to simple truths about the local environment. He suggests that "mindreading is typically used for matters of actual or potential interest to the agents concerned", and that on such points, "it is typically easier to work down from an initial hypothesis of total knowledge than to work up from an initial hypothesis of total ignorance" (2024, 45). However, if others are by default taken to be knowledgeable on all points of actual or potential interest, one might worry about the fit between this default setting and our evident drive to tell each other things in conversation.[19]

Here it matters that what we instinctively register as commonly known in the waiting room is the fact of the loudspeaker announcement, as opposed to, say, the fact that the carpet has a pattern of grey and blue dots, even if both of these details are perceptually available to everyone. Loudspeakers reliably capture attention, puncturing the silence of the waiting room with content of a type we expect to be relevant to action. For those with no particular expectations about carpet patterns—which is to say most of us—a run-of-the-mill carpet pattern is unsurprising and unlikely to command attention. None of us needs to draw the others' attention to the loudspeaker announcement, or tell them about it, if we are sitting in the waiting room together; we know from past experience that this is the kind of stimulus that others will register as we do. We learn such regularities about the individual knowledge of others through dyadic interaction with them, starting with semi-random pointing in infancy: sometimes, when you point something out to someone, that target comes as a surprise to them (*oh*) and sometimes it does not. In infant-caregiver dyads, the sheer fact that the infant is successfully pointing to something is enough to trigger surprise in the caregiver, and typically a naming response (Olson & Masur, 2011). The naming response simultaneously confirms joint attention to the object and yields surprise for the infant learner, whose emergent linguistic abilities generate some expectations, given the regularity of the lexicon, with some genuine prospect for surprise, given the infant's merely partial command of language. Infants do not point out objects to each other, presumably because doing so is not informationally or otherwise rewarding (Ninio, 2016). Meanwhile, infants who keep pointing out the same thing to a caregiver will also experience diminishing rewards from doing so; there is more reward in finding something fresh, on which the dyad has not previously interacted. The learned value of distinguishing what is fresh for the dyad explains our early competence in tracking personal common ground; indeed, by 12-18 months of age, infants have learned that adults also point out what is fresh to a dyad. Amid toys that were all

[19] One might have a similar worry about Robert Gordon's (2021) agent-neutral coding theory: to the extent that our grasp of the world is by default undifferentiated as to whose representations are whose, we face challenges in explaining the motivations behind intentional communication.

equally familiar to the infant, and after playing with all but one of these toys with a particular adult, infants interpreted this adult's vague pointing gesture (coupled with "Oh wow! Look at that!") as directed towards the one toy the pair had never engaged with together (Tomasello & Haberl, 2003).

By three years of age, children advance beyond tracking personal common ground to tracking cultural common ground: when meeting someone new from one's ingroup, a child expects this ingroup stranger to find culturally well-known facts unsurprising, as if the child and stranger had already interacted on the point (Liebal, Carpenter, & Tomasello, 2013). The broadly interactive dynamics of cultural common ground make it unnecessary to have interacted with any particular in-group stranger in order to have something like dyadic knowledge with them already on a point in cultural common ground: one can know a new individual will not be surprised by this fact by generalizing from past interactions with others.

Tomasello himself takes the positive epistemic quality of communication for granted, or rather ascribes it to an innate prosocial tendency to be informative and cooperative, which raises puzzles about why infants do not point things out to each other. Tomasello sees nine-month-old infants as "able to form with [adult] others a joint agent, comprising two partners who act and experience things together" (2019, 87); however, his explanation of the structure of this joint agency invokes a recursive structure of thinking about the other party's communicative intentions, of a type that he insists the young infant cannot literally be engaging in (2019, 56), making it somewhat mysterious just what is going on.

The knowledge-first approach, by contrast, explains the structure and existence of communication by its capacity to generate knowledge, supporting successful action, and invoking just simple mechanisms of extrinsic and intrinsic reward-driven learning. Conversation is driven mainly by curiosity, with surprise as reward. Points on which we are all knowledgeable will no longer yield surprise for any of us, but as common knowledge accumulates, we learn gradually more sophisticated ways of finding the world's remaining pockets of surprise, including ways that involve exploring and exploiting the knowledge of others. The steadily expanding default zone in which we are all knowledgeable is built together with a steadily keener sense for what is unknown: common knowledge expands in large measure because curiosity attracts us to the edge-case zones where knowledge is limited. Our everyday practical purposes are well served by the extent to which the world lies open to view, but our curiosity is engaged by the extent to which the world also lies open to discovery. As long as we have natural curiosity, we will be motivated to explore the world actively, while learning from experience that more effective discovery is enabled through communication with others. We may or may not be surprised to discover that a knowledge-first approach to communication can start to reveal

how common knowledge could be possible without recursion, and within the bounds of our psychological capacities. Arguably, this is just the kind of development we should already have started to expect from the first sentence of *Knowledge and its Limits:* "Knowledge and action are the central relations between mind and world" (Williamson, 2000, 1).[20]

References:

Ameka, F. (1992). Interjections: The universal yet neglected part of speech. *Journal of Pragmatics, 18*(2-3), 101-118.

Aumann, R. J. (1976). Agreeing to Disagree. *Annals of Statistics, 4*(6), 1236-1239.

Bateson, M. C. (1975). Mother-infant exchanges: the epigenesis of conversational interaction. *Annals of the New York Academy of sciences, 263*(1), 101-113.

Bavelas, J. B. (2022). *Face-to-face dialogue: Theory, research, and applications*: Oxford University Press.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology, 79*(6), 941.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of communication, 52*(3), 566-580.

Bavelas, J. B., Gerwing, J., & Healing, S. (2017). Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *Journal of pragmatics, 121*, 91-112.

Berke, M. D., Walter-Terrill, R., Jara-Ettinger, J., & Scholl, B. J. (2022). Flexible goals require that inflexible perceptual systems produce veridical representations: Implications for realism as revealed by evolutionary simulations. *Cognitive Science, 46*(10), e13195.

Binmore, K., & Samuelson, L. (2001). Coordinated action in the electronic mail game. *Games and Economic Behavior, 35*(1-2), 6-30.

Christoffels, I., & De Groot, A. (2005). Simultaneous interpreting. In J. F. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 454-479). Oxford: Oxford University Press.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62-81.

Crawford, V. P. (2019). Experiments on cognition, communication, coordination, and cooperation in relationships. *Annual Review of Economics, 11*(1), 167-191.

Dessalles, J.-L. (2020). Language: The missing selection pressure. *Theoria et Historia Scientiarum, XVII*, 7-57.

Dingemanse, M. (2023). Interjections. In E. van Lier (Ed.), *Oxford Handbook of Word Classes* (pp. 477-492). Oxford: Oxford University Press.

Dingemanse, M. (2024). Interjections at the Heart of Language. *Annual Review of Linguistics, 10*, 257-277.

Dingemanse, M., & Enfield, N. (2024). Interactive repair and the foundations of language. *Trends in cognitive sciences*.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., . . . Manrique, E. (2015). Universal principles in the repair of communication problems. *PloS one, 10*(9), e0136100.

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PloS one, 8*(11), e78273.

Dunbar, R., Duncan, N. D., & Nettle, D. (1995). Size and structure of freely forming conversational groups. *Human Nature, 6*, 67-78.

Dunbar, R., Marriott, A., & Duncan, N. D. C. (1997). Human conversational behavior. *Human Nature, 8*(3), 231-246.

Emler, N. (1994). Gossip, reputation, and social adaptation.

Enfield, N. J. (2017). *How we talk: The inner workings of conversation*: Basic Books.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Gardner, R. (2001). *When Listeners Talk*. Amsterdam: John Benjamins.

[20] For helpful discussion, I would like to thank Sara Aronowitz, David Barnett, Nilanjan Das, Jan Engelmann, Julia Haas, Liang-Zhou Koh, Andrew Lee, Johannes Mahr, Mohan Matthen, Bence Nanay, Paulina Sliwa, Leo Tenenbaum, Sergio Tenenbaum, Timothy Williamson, Snow Zhang, and audiences at the Lauener Prize Symposium in Bern, Switzerland, the University of Vienna, the University of Antwerp, and Columbia University. Thanks also to the Social Sciences and Humanities Research Council of Canada for funding my research.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual Review of Psychology, 68*(1), 101-128.
Greco, D. (2014). Could KK be OK? *The Journal of Philosophy, 111*(4), 169-197.
Greco, D. (2015). Iteration principles in epistemology I: Arguments for. *Philosophy Compass, 10*(11), 754-764.
Guellaï, B., Hausberger, M., Chopin, A., & Streri, A. (2020). Premises of social cognition: Newborns are sensitive to a direct versus a faraway gaze. *Scientific reports, 10*(1), 9796.
Guthrie, A. M. (1997). On the systematic deployment of okay and mmhmm in academic advising sessions. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 7*(3), 397-415.
Haas, J. (2022). Reinforcement learning: A brief guide for philosophers of mind. *Philosophy Compass*. doi:10.1111/phc3.12865
Haviland, J. B. (1977). *Gossip, reputation, and knowledge in Zinacantan*: University of Chicago Press Chicago.
Heal, J. (1978). Common knowledge. *The Philosophical Quarterly (1950-), 28*(111), 116-131.
Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Action* (pp. 299-345). Cambridge: Cambridge University Press.
Heritage, J. (2012). Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction, 45*(1), 1-29.
Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one, 10*(8), e0136905.
Jaffe, J. (1987). Parliamentary procedure and the brain. In A. W. Siegman & S. Feldstein (Eds.), *Nonverbal behavior and communication* (pp. 21-33). Hillsdale, NJ: Psychology Press.
Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T.-y., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20*(8), 963-973.
Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.
Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *Journal of human evolution, 40*(5), 419-435.
Koivisto, A. (2016). Receipting information as newsworthy vs. responding to redirection: Finnish news particles aijaa and aha (a). *Journal of pragmatics, 104*, 163-179.
Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of personality and social psychology, 4*(3), 343.
Lederman, H. (2018). Uncommon knowledge. *Mind, 127*(508), 1069-1105.
Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences, 20*(1), 6-14.
Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.
Lewis, D. (1978). Truth in fiction. *American Philosophical Quarterly, 15*(1), 37-46.
Lewis, P. A., Birch, A., Hall, A., & Dunbar, R. I. (2017). Higher order intentionality tasks are cognitively more demanding. *Social cognitive and affective neuroscience, 12*(7), 1063-1071.
Liebal, K., Carpenter, M., & Tomasello, M. (2013). Young children's understanding of cultural common ground. *British Journal of Developmental Psychology, 31*(1), 88-96.
Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & De Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science, 36*(4), 698-713.
Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience, 9*(11), 1432-1438.
Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science, 317*(5834), 82-82.
Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in cognitive sciences, 20*(9), 689-700.
Nagel, J. (2024). Natural Curiosity. In A. Logins & J.-H. Vollet (Eds.), *Putting Knowledge to Work: New Directions in Knowledge-First Epistemology* (pp. 170-200). Oxford: Oxford University Press.
Ninio, A. (2016). Bids for joint attention by parent–child dyads and by dyads of young peers in interaction. *Journal of child language, 43*(1), 135-156.
Ninio, A. (2018). *Pragmatic development*: Routledge.
Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of pragmatics, 41*(5), 866-891.
Olson, J., & Masur, E. F. (2011). Infants' gestures influence mothers' provision of object, action and internal state labels. *Journal of child language, 38*(5), 1028-1054.
Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., . . . Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science advances, 9*(13), eadf3197.
Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

Shepherd, S. V. (2010). Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience, 4*, 5.
Stalnaker, R. (2014). *Context*: OUP Oxford.
Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience, 15*(8), 520-535.
Stivers, T. (2021). Is conversation built for two? The partitioning of social interaction. *Research on Language and Social Interaction, 54*(1), 1-19.
Tolins, J., & Fox Tree, J. E. (2014). Addressee backchannels steer narrative development. *Journal of pragmatics, 70*, 152-164.
Tomasello, M. (2019). *Becoming human: A theory of ontogeny*: Belknap Press.
Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental Psychology, 39*(5), 906.
Vanderschraaf, P., & Sillari, G. (2022). Common Knowledge. *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/common-knowledge/
Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition, 14*(1), 129-182.
Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition, 210*, 1-15.
Williamson, T. (2000). *Knowledge and its Limits*. New York: Oxford University Press.
Williamson, T. (2024). Where did it come from? Where will it go? In A. Logins & J.-H. Vollet (Eds.), *Putting Knowledge to Work: New Directions for Knowledge-First Epistemology* (pp. 21-70). Oxford: Oxford University Press.
Wohltjen, S., & Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences, 118*(37), e2106645118.
Wu, R.-J. R., & Heritage, J. (2017). Particles and epistemics: Convergences and divergences between English and Mandarin. In *Enabling human conduct* (pp. 273-298): John Benjamins.
Yngve, V. H. (1970). *On getting a word in edgewise.* Paper presented at the Chicago Linguistics Society, 6th Meeting, 1970.
Yoo, H., Bowman, D. A., & Oller, D. K. (2018). The origin of protoconversation: An examination of caregiver responses to cry and speech-like vocalizations. *Frontiers in psychology, 9*, 1510.
Zuckerman, C. H., & Enfield, N. (2023). The limits of thematization. *Journal of Linguistic Anthropology, 33*(3), 234-263.