SOCIAL SCIENCE

The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation

Andrew Reece[1]*†, Gus Cooney[2]*†, Peter Bull[3], Christine Chung[3], Bryn Dawson[1], Casey Fitzpatrick[3], Tamara Glazer[3], Dean Knox[2], Alex Liebscher[1], Sebastian Marin[1]

People spend a substantial portion of their lives engaged in conversation, and yet, our scientific understanding of conversation is still in its infancy. Here, we introduce a large, novel, and multimodal corpus of 1656 conversations recorded in spoken English. This 7+ million word, 850-hour corpus totals more than 1 terabyte of audio, video, and transcripts, with moment-to-moment measures of vocal, facial, and semantic expression, together with an extensive survey of speakers' postconversation reflections. By taking advantage of the considerable scope of the corpus, we explore many examples of how this large-scale public dataset may catalyze future research, particularly across disciplinary boundaries, as scholars from a variety of fields appear increasingly interested in the study of conversation.

INTRODUCTION

Conversation hardly needs introduction. It is a uniquely human act of cooperation that requires exquisite coordination across many levels of cognition (*1*–*4*). It is the seat of language acquisition (*5*). Its turn-taking system emerges early in development (*6*, *7*) and shows parallels in nonhuman primates and other animals (*8*, *9*). It is how group members absorb and transmit culture (*10*, *11*). It is the primary tool that humans use to form and maintain their social relationships (*12*, *13*). It has a substantial impact on people's mental and physical health (*14*, *15*), and more recently, generative models of conversation have emerged as a major milestone in artificial intelligence (*16*–*18*).

Despite its centrality, conversation's complexity has hampered its empirical study: Conversation is characterized by a strong degree of interdependence between speaking partners, in which one's words and behavior are adjusted rapidly in response to what one's partner is doing; conversation is staggeringly multimodal, involving information transmission across linguistic, paralinguistic, and visual channels simultaneously; and last, conversation is highly contextualized, in which people play certain social roles, pursue specific goals, and negotiate status and power hierarchies. In turn, this complexity presents numerous scientific challenges, from operationalization to measurement to statistical modeling. However, here, we demonstrate that recent technological advances have begun to offer solutions to these challenges, placing previously inaccessible research questions within reach and offering considerable opportunity for interdisciplinary collaboration.

Historically, progress on conversation research has been catalyzed by large public datasets, such as the Map Task Corpus (*19*), the Switchboard Corpus (*20*), or newer multimodal datasets, such as the MELD (*21*, *22*) and OMG-Empathy datasets (*23*) [for a review, see (*24*)]. While these datasets have advanced conversation science, none includes a large sample of naturalistic conversation, with full audio and video recordings, together with speakers' detailed postconversation reports.

We collected such a dataset of 1656 unscripted conversations over video chat that comprise more than 7 million words and 850+ hours of audio and video. Overall, our corpus includes more than 1 terabyte of raw and processed recordings. The corpus draws on a large and diverse sample of participants, aged 19 to 66, from all over the United States. Participants were paired using an automatic matching algorithm of our own design and were simply instructed to have a conversation with one another for at least 25 min, although many talked for much longer. The conversations occurred during 2020 and, thus, offer a unique perspective on one of the most tumultuous years in recent history, including the onset of a global pandemic and a hotly contested presidential election. The corpus is among the largest multimodal datasets of naturalistic conversation, which we refer to collectively as the CANDOR corpus (Conversation: A Naturalistic Dataset of Online Recordings).

Large amounts of raw data alone are not sufficient to advance the study of conversation. In other domains, growth in computational power, the use of crowdsourcing platforms, and technological advances in machine learning, e.g., language and signal-processing algorithms such as Word2Vec, BERT, and ResNet, have proven to be yet another catalyst of scientific advancement, enabling discovery and inference at scale (*25*–*29*). In this spirit, we applied an elaborate computational pipeline to quantify features of conversation such as overlaps and pauses, second-by-second variation in facial features, and full transcripts with accompanying prosodic characteristics of speech. Last, we collected a battery of psychological measures from the participants, including trait-level measures such as personality, as well as people's opinions about their conversation partner and their feelings about the overall conversation.

We explore the corpus in five sections. First, we use the corpus to replicate key findings from the literature on turn-taking. In doing so, we developed algorithmic procedures to segment speech into conversational turns—a preprocessing step necessary to study conversation at scale—and demonstrate how analytic results hinge critically on the choice of appropriate segmentation algorithms. Second, we explore the relation between conversation and

[1]BetterUp Inc., San Francisco, CA 94103, USA. [2]University of Pennsylvania, Philadelphia, PA 19104, USA. [3]DrivenData Inc., Berkeley, CA, 94709, USA.
*Corresponding author. Email: andrew.reece@betterup.com (A.R.); guscooney@gmail.com (G.C.)
†These authors contributed equally to this work.

psychological well-being. Third, we connect patterns of turn-taking to people's well-being, exploring the nuanced relationship between the structure of conversation and its psychology. Fourth, we apply a series of computational models to text, audio, and visual data and extract detailed measures of turn-by-turn behavior that we use to investigate an open question in the literature: What distinguishes a good conversationalist? Last, we briefly overview our corpus's topical, relational, and demographic diversity using a mixed-methods approach: We analyze the way the national discourse shifted during a tumultuous year, review the entire corpus qualitatively to identify conversations that were particularly high in rapport, and examine the ways people alter their speech, listening patterns, and facial expressions when they talk to partners from diverse backgrounds and identity groups.

To structure these findings, we articulate a framework based on a more "vertically integrated" approach to the study of conversation. Our hierarchy of conversation spans (i) "low-level" mechanical features of conversation, such as turn-taking, which delineate the structure of interaction; (ii) "mid-level" information streams, such as semantic exchange, psycholinguistic markers, and emotion expressions, which represent the subjective content of turn-by-turn conversation; and ultimately, (iii) "high-level" judgments reported after conversation, such as people's enjoyment and the impressions they formed of their conversation partners. Many of our results demonstrate that these levels cannot be meaningfully studied in isolation.

The findings we present from the corpus are far from exhaustive. Rather, they are intended as a launching point for future research and collaboration. In other contexts, the emergence of grassroots consortiums [e.g., (*30*, *31*)] has allowed scientists to pool ideas and resources, using "big" science to address large unanswered questions (*32*). Many of our results demonstrate not only the considerable advantages associated with studying conversation through a multidisciplinary lens but also the need for larger collaborative efforts to make empirical progress. Our goal is that the corpus will help build an interdisciplinary science of conversation.

Many disciplines have been drawn to the study of conversation, some for decades (e.g., conversation analysis, sociology, communications, pragmatics, and psycholinguistics) and some more recently (e.g., cognitive and social psychology, neuroscience, organizational behavior, political science, computational linguistics, natural language processing, and artificial intelligence). In all cases, it appears that progress has been catalyzed by rich datasets, new analytic frameworks, and empirical findings that beckon further collaboration, all of which we have intended to provide. Together, we hope that these offerings will advance the study of the most fundamental of all human social activities: the spoken conversation.

CORPUS CONSTRUCTION

Between January and November 2020, six rounds of data collection yielded a total of 1656 dyadic conversations that were recorded over video chat (see table S1). In what follows, we explain our recruitment method and the construction of the final dataset.

Methods

Recruitment

Initial survey. Our target population consisted of people 18+ years of age who live in the United States. We recruited participants using Prolific, an online crowdwork platform. This study was approved by Ethical & Independent Review Services, protocol #19160-01.

Before entering the study, candidate participants were asked to read a consent form that explained the following: (i) They would have a conversation with another individual that would last at least 25 min; (ii) audio and video would be recorded; (iii) they would complete a series of surveys before and after the conversation; (iv) they would be paid $0.85 for completing the initial survey and an additional $14.15 upon full completion of the recorded conversation and postconversation survey; (v) their data, including the video and audio recordings, would be shared with other researchers and could be made publicly available; and (vi) participation carried a risk of personal identification because of the audio and video recordings. Because of the sensitive nature of releasing personally identifiable recordings and the study design that required participants to meet for a video call after they gave initial informed consent, we then asked participants to verify again that they were comfortable having a recorded conversation with a stranger. Only candidates who both indicated and reaffirmed their consent were permitted to continue as study participants.

Participants were then asked a series of questions to determine their availability during the following week. After doing so, they received a follow-up email within 24 hours.

In some data collection rounds, participants filled out a small number of additional psychological measures (see the round variable in the Data Dictionary for details). Last, they were given instructions to submit a request for compensation.

Matching. A matching procedure was carried out once daily based on the participants' stated availability during the next week. Unmatched participants with overlapping availability were paired. No demographic information was used in the matching process. Once matched, participants were notified by email of the time and date of their conversation. A second email that contained a link to a survey, which guided participants through the next phase of the study, was sent 1 hour before the scheduled conversation.

Participants

Of the participants who completed the intake survey (approximately $N = 15{,}000$), approximately 3500 were matched with another participant, returned to have a conversation, and provided audio and video recordings that we were able to process automatically with our pipeline. Naturally, given the difficulties inherent in scheduling strangers to meet in a video chat room on the internet at a specific date and time, we experienced cancelations, no-shows, technologically confused participants, and other obstacles over the course of data collection. For example, slightly more than 3000 participants reported at least one instance in which their partner simply did not arrive for the scheduled video call. We addressed this contingency by compensating participants who experienced no-shows with $1.50 and offering them the opportunity to rejoin our matching pool the next day.

All told, we recorded approximately 2000 completed conversations by the end of the data collection period that totaled around 1000 hours of footage. An additional human review of all of our conversations flagged approximately 300 conversations for removal. Conversations were eliminated for two main reasons: Another individual appeared on camera who did not consent to be filmed (e.g., a participant wanted their conversation partner to

say hello to one of their children) or technical issues made the audio or video recording unusable.

Our final dataset included 1656 conversations and 1456 unique participants who spanned a broad range of gender, educational, ethnic, and generational identities (see table S2). "Unique participants" refer to the number of participants who had at least one conversation, as more than 50% of our sample had more than two conversations and 33% had more than three conversations. Participants who had multiple conversations did not do so back-to-back and, in most cases, held several conversations distributed across the data collection period (see table S2).

Preconversation survey

A preconversation survey measured participants' current mood (i.e., valence and arousal; see the Data Dictionary). The survey then reminded participants to turn on their webcam and microphone, to ensure that their conversation lasted at least 25 min, and to return to the survey tab in their browser after the conversation was over to complete the postconversation survey. Last, a link was provided to the video chat room.

The conversation

Clicking the "Join Conversation" link opened a new video chat window. Recording began as soon as the first conversation partner joined. Participants were asked to wait at least 5 min for their partner to arrive. Sessions in which only one participant arrived were discarded.

With respect to the conversation content, participants were not given specific instructions—they were simply told to "talk about whatever you like, just imagine you have met someone at a social event and you're getting to know each other." Then, the participants were instructed to have a conversation for at least 25 min, although their duration varied considerably (mean length = 31.3 min, SD = 7.96, minimum = 20 min).

Conversations were recorded digitally using a web application based on the TokBox OpenTok Video API and were conducted via camera-connected displays and microphones. Most conversations were conducted computer to computer, but a mobile device was used occasionally. Upon completion, the participants ended the recording session and returned to complete the postconversation survey.

Postconversation survey

Participants were asked first whether any issues prevented them from completing the conversation. If so, then we offered them the opportunity to reschedule with a new partner by responding with their updated availability. Otherwise, participants went on to complete a postconversation survey, in which they reported their perceptions of their conversation partners, their feelings about the overall conversation, their personality, and so forth. For details, see the Data Dictionary (the link is in the Data and materials availability statement), which describes all measures. Last, they were thanked for their participation and provided instructions on how to submit a request for payment.

Available data

The two primary outputs from the conversation collection process were the survey responses and the video archive that contained the videos and metadata for each conversation.

Survey. The survey data consisted of (i) the initial survey administered during the screening stage, (ii) the preconversation survey, and (iii) the postconversation survey. The survey responses were processed via the Qualtrics API into a flat file of comma-separated values, and participants' responses were recorded at the conversation level. For full details about the survey items, please refer to the Data Dictionary.

Data processing and feature extraction. Please see the Supplementary Materials for a detailed explanation of the way the video and audio recordings were processed into unified, analysis-ready formats. Briefly, across all modalities—textual, acoustic, and visual—our goal was to extract and streamline as much information as time and technology would permit, producing a user-friendly corpus for researchers to use and improve upon. The resulting feature sets extracted from our processing pipeline are described in full in the Data Dictionary.

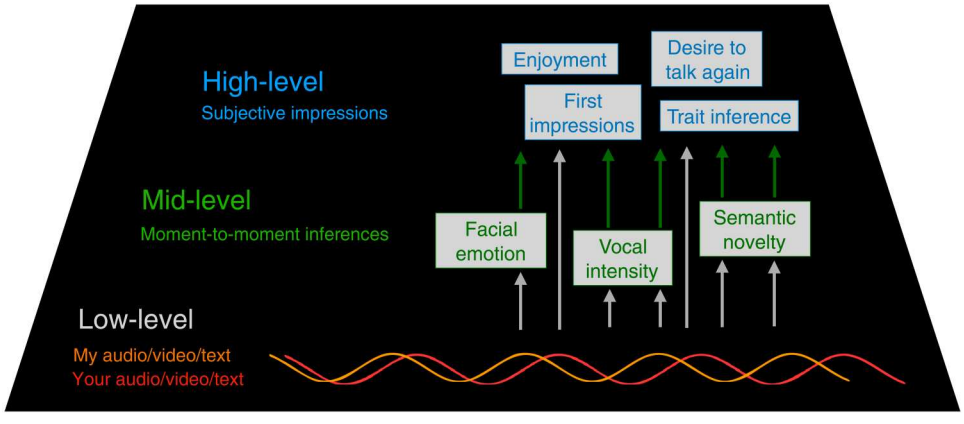

Fig. 1. A framework for studying conversation. The results are organized according to an analytic framework that distinguishes between three related levels of conversation. Low-level features can be observed directly, vary over short time periods, and often relate to conversational structure (e.g., a pause at the end of a speaker's turn). Mid-level features are generally inferred indirectly by human perceivers or algorithms that approximate human perception, vary on a medium-frequency or turn-by-turn basis, and capture linguistic or paralinguistic conversational content (e.g., a happy facial expression or vocal emotional intensity). High-level features relate to people's subjective judgments of a conversation (e.g., postconversation-reported enjoyment or people's evaluations of their partner). Subsequent sections present empirical results at each level of the hierarchy, as well as analyses that demonstrate the interplay across levels that we believe will represent an increasingly common and important type of research.

With future advances in preprocessing and machine learning algorithms, human review, and additional effort, we anticipate that current and future scholars will obtain considerable additional value from this corpus. We encourage researchers who develop improvements to this corpus to make their improvements publicly available to the broader scientific community.

CORPUS FRAMEWORK

To guide our exploration of the CANDOR corpus, we embraced three principles: (i) Conversation is constructed around a highly cooperative system of turn-taking; (ii) understanding the full complexity of conversation requires insights from a variety of disciplines that, although they examine the same phenomenon, often remain siloed in their research questions and analytic tools; and (iii) examining conversation computationally and at scale is an enduring challenge, but new technologies, particularly advances in machine learning, promise to unlock many aspects of conversation that were previously inaccessible to empirical research.

The framework

We propose an organizing framework that classifies conversational features as low-level, mid-level, and high-level (see Fig. 1). Low-level features are closest to the raw signals in the audio, video, and text of a conversation recording and often vary on a nearly continuous time scale. Although some degree of inference is often necessary to generate even these features, such as extracting vocal markers from processed audio signals or the linguistic inferences made by an automated transcription service, these outputs are sufficiently concrete and specific (e.g., pitch, turn duration, eye gaze, etc.) to constitute the objective properties of conversation from which higher-order inferences are derived.

Next, high-level features are individuals′ subjective judgments about their conversations, formed on a coarse time scale and reflected in the postconversation survey responses. Survey items included measures of liking, enjoyment, and conversational flow, as well as evaluations of one′s partner′s social status, intelligence, and personality. The value of these postconversation ratings is considerable, as they allow in-conversation behaviors and post-conversation impressions to be linked.

Between these levels, we identified numerous features related to subjective perceptions of the interaction that typically vary on an intermediate time scale. These mid-level features capture intraconversational psychology and are usually computed using a suite of algorithmic tools that were trained to attend to specific aspects of speech, sound, and movement to infer a psychological content, such as a happy facial expression, an increasing intensity in one′s voice, and a timely change of subject: Noticing these conversational moments requires a mix of sense and sense-making—whether by human or machine—and is analytically distinct from low- and high-level phenomena.

We refer interested readers to the Supplementary Materials for additional theoretical implications of this tiered framework to study conversation. A simple example demonstrates the nature of low-, mid-, and high-level features: The contraction of a person′s zygomaticus major muscle is, in principle, an observable (low-level) feature. Most people recognize this contraction pattern as a smile, a momentary expression of happiness (mid-level inference). Last, individuals who smile frequently during a conversation may also report that they had an enjoyable experience overall (a high-level, subjective report). We used this framework as a heuristic to organize a diverse array of findings across the rich dataset.

First, in what follows, we present results related to conversational mechanics, the lowest level of our framework that covers features such as turn exchange and backchannel feedback, and the fact that studying these low-level mechanical features requires developments in transcript segmentation. Second, at the high level, we examine the way conversation influences an individual′s well-being. Third, we demonstrate the interplay between levels and show, for example, the way an individual′s low-level speed of turn exchange relates to their partner′s high-level enjoyment of the conversation. Fourth, we explore the middle layer of the corpus by extracting psychologically rich features with an array of computational models. These fine-grained measures of turn-by-turn interaction are used to link the middle and the high level and to investigate a basic unanswered question in conversation research: What constitutes a good conversationalist? Last, we end with a mixed-method report that explores our corpus′s topical, relational, and demographic diversity.

RESULTS

The turn-taking system and algorithms for transcript segmentation

A hallmark of conversation is that there is no predetermined order of who should speak, about what, and for how long. Given this precarious starting point, it is something of a marvel that conversation works so well. Managing such variability depends upon a complex, highly coordinated system of turn-taking.

The turn-taking system has many elements, but three basic components are the following: (i) turn exchange—the way people manage to pass the floor back and forth in an orderly and efficient manner; (ii) turn duration—how long speakers talk before they turn over the floor; and (iii) backchannel feedback—the active engagement that listeners display while speakers are talking, such as the use of nods or short utterances—"mhm," "yeah," and "exactly"—to convey understanding and encouragement. The scope of our data permits close investigation of these basic features of conversation. We do so with two primary objectives: first, to demonstrate that key results from the literature are replicated in the corpus and, second, to explore ideas related to studying conversation at scale. Our ideas related to this revolve around the thorny question of what constitutes a turn and the combined problems that there is no agreed upon definition, as well as the difficultly of implementing such definitions computationally.

Turn exchange

One important finding related to turn exchange is that the average interval between turns (as well as the median and modal response) is a short gap of approximately 200 ms, approximately the duration of an eyeblink, which is a figure that appears to be consistent across languages and cultures (*33*, *34*). Here, we replicated this previous work in our large corpus of video-recorded conversations. Following Heldner and Edlund (*35*), we applied a procedure to classify communication states in our entire corpus to obtain a time series in which the presence or absence of speech for both speakers is recorded at 10-ms intervals. This allowed us to identify within- and between-speaker intervals with high temporal precision. The resulting taxonomy included gaps (between-speaker silences), pauses

(within-speaker silences), overlap (between-speaker overlap), and within-speaker overlap (when one speaker begins to speak while their conversation partner is still speaking, such as in the case of an attempted interruption or backchannel). Here, we focus on gaps and overlaps.

Figure 2 shows that gaps and overlaps followed an approximately normal distribution centered on zero (equivalent to a perfectly timed, no-gap, no-overlap speaker transition). Gaps and overlaps were overwhelmingly less than 1-s long and often much shorter, with a median of 380 ms and −410 ms, respectively. Gaps represented approximately half (52.1%) of all speaker transitions, while slightly less than half of speaker transitions were overlaps (47.9%). Overall, the median between-speaker interval was a fleeting 80 ms.

The brief interval between turns is particularly notable because the length of these intervals is much shorter than the time it takes a person to react and produce a spontaneous utterance. This means that listeners must be predicting the end of a speaker′s turn in advance. Previous research has identified mechanisms by which people accomplish this feat, for example, by using various syntax and prosody cues [(*36*–*38*); for a review, see (*39*)].

Overall, these figures closely match those in the previous literature (*35*, *40*) and replicate earlier findings in a dataset with nearly half a million speaker transitions. We also replicate this previous literature in the increasingly important domain of video-mediated communication—it appears that basic conversational dynamics during such conversation closely resembled those seen in face-to-face interaction [c.f., (*41*)].

Last, note that the distribution of turn intervals observed depends upon the definition of a turn being implemented. For example, an automated corpus work tends to define a "turn" as any stretch of speech before the speakers switch. This definition is based on timing without consideration of the content of the turn. Recent work has emphasized the way the conclusions drawn about the timing of turns depends critically upon these definitional choices (*34*, *42*). Undoubtedly, future work will examine this question in more detail, and in relation, in the following sections, we explore different algorithms to segment turns at scale. We also return to the dynamics of turn exchange later to examine the relation between low-level conversational features and higher-level psychological outcomes.

Turn duration

Turn duration, or how long people hold the floor, is another basic feature of the turn-taking process. Compared to the sizable literature on turn exchange, turn duration remains relatively understudied, attributable, in part, to a lack of available datasets of suitable quality and scale. Here, we show how advancing the empirical study turn duration will require new algorithms to segment transcripts into turns.

Talk ratio. Speaker turn duration can be conceptualized in two ways. The first is to measure the time that a given speaker holds the floor overall (e.g., speaker A talked for 15 min of a 20-min

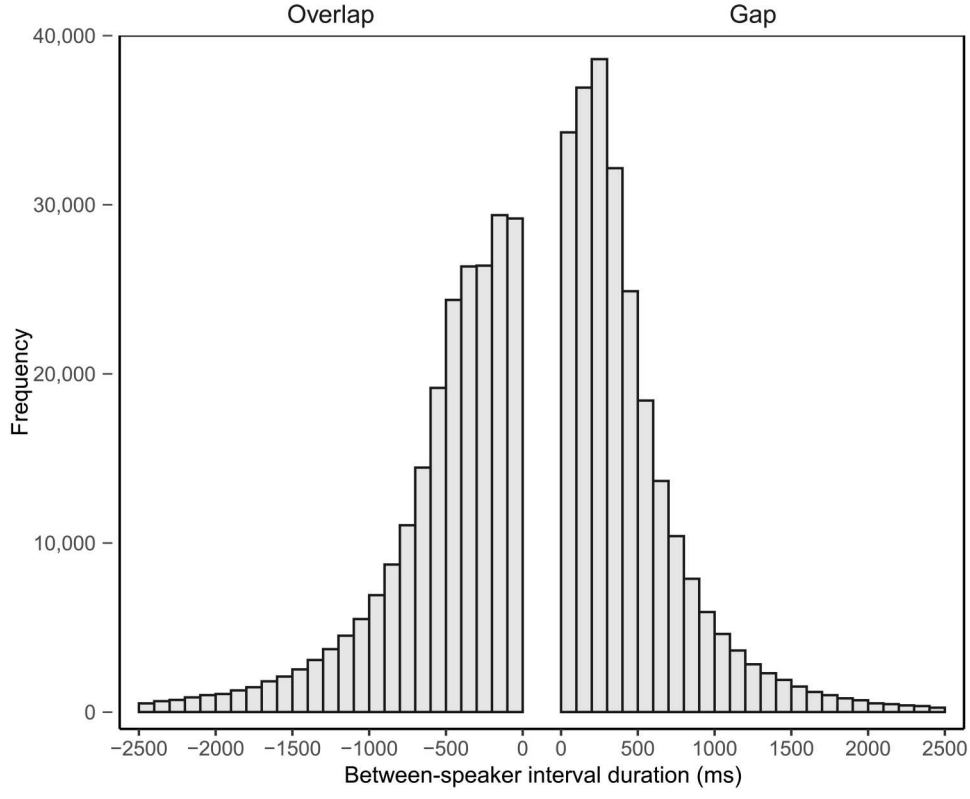

Fig. 2. Distribution of gaps and overlaps across speaker transitions. Negative intervals are classified as "overlaps," indicating the presence of simultaneous speech on the part of two conversation partners. Positive values are "gaps," indicating a stretch of silence between turns. The results indicated that the median between-speaker turn interval was +80 ms and was distributed approximately normally. These results are similar to those previously observed in conversations across many cultures and communication modalities (*33*). Short gaps are consistently the most common type of turn transition in naturally occurring conversation.

conversation, or 75% of the time, while speaker B talked for only 5 min, or 25% of the time). The second, more fine-grained approach is to measure the duration of individual turns.

The individual turn approach offers a richer portrait of conversation. For example, people's overall talk times necessarily fail to distinguish between two people exchanging extended monologs versus two people engaging in high-speed back-and-forth conversation. Finer-grained observation at the turn level has the potential to reshape literatures that currently rely on coarser overall measurements, such as studies that have examined the relation between talk time overall and perceptions of leadership, likability, and dominance (*43*, *44*). However, progress in this direction makes it necessary to grapple first with a difficult definitional dilemma at the heart of large-scale conversation research: What do we mean by a turn?

Individual turn duration. To date, the most rigorous definition of conversational turns is found in the conversation analysis literature (and that of related disciplines), in which researchers have developed rubrics to transcribe conversations manually [e.g., (*45*)]. However, large corpora such as ours, which include hundreds of thousands of potential turn boundaries, make conventional manual coding highly impractical. Scaling these efforts computationally would require a precisely defined series of steps—i.e., an algorithm—to be constructed to organize two speakers' streams of speech into psychologically meaningful turns. Below, we demonstrate three candidate algorithms, or "turn models," and describe their design rationale and limitations. We consider these algorithms useful, if early, starting points for continued research on the subject of conversational turns and, by extension, turn duration.

Our most basic turn model, Audiophile, formalizes a simple assumption: A turn is what one speaker says until their partner speaks, at which point the partner's turn begins, and so on. This is essentially the way the AWS Transcribe API parses recorded speech into turns. While useful as an easy-to-implement benchmark, Audiophile had some salient drawbacks. For example, because of brief sounds (such as laughter) or short bits of cross-talk, Audiophile often broke up turns too aggressively and created several small turns that virtually any human observer would have regarded, syntactically and psychologically, as a single turn. However, as we demonstrate below, approximating "psychologically real" turns is a difficult task that has seemingly stymied research on turn duration at scale.

In this section and the next, we describe the development of two competing models that sought to approximate human's perceptions of turns better. We begin with Cliffhanger, which attempts to capture the duration of people's turns more accurately by segmenting turns based on terminal punctuation marks (periods, question marks, and exclamation points), as generated by automated transcription. From Cliffhanger's "perspective," once the transcript indicates that speaker A has begun speaking, if speaker B interjects during A's sentence (i.e., before A reaches a terminal punctuation mark), then B's utterance is shifted into a new turn after A's sentence concludes. Subsequent sentences on A's part are then assigned to a new turn that succeeds B's interruption. In essence, Cliffhanger disallows turn exchanges until after the primary speaker has finished their current sentence.

The simple example in Fig. 3 makes the Cliffhanger procedure clear. As depicted, the mean and median Cliffhanger turn durations were four and five times greater, respectively, compared to that of Audiophile. The clear implication is that studies of turn duration will rely heavily on the researcher's choice of turn model. How might we empirically determine whether Cliffhanger is a more suitable model for studying turn duration?

One indirect test of a turn model's performance is its face validity with respect to producing turns that are consistent with human intuition. For example, anecdotal evidence suggests that turn length is correlated with a conversation's quality: People who enjoy a conversation will likely have more to say and, thus, may be expected to take longer turns on average. We did observe this relation between speakers' turn duration and their enjoyment of the conversation overall—but only when we used turns Cliffhanger generated, not Audiophile. For Audiophile turns, the association between people's median turn duration and their enjoyment was not significant [b = 0.07, 95% confidence interval (CI) = [−0.04, 0.18], $t(3255) = 1.32$, $P = 0.19$]. By contrast, for Cliffhanger turns, the relationship between median turn duration and people's enjoyment was significant [b = 0.06, 95% CI = [0.03, 0.09], $t(3255) = 4.12$, $P < 0.001$]. This analysis illustrates the way results depend critically upon turn model selection and, in doing so, introduces an important "researcher degree of freedom" (*46*). This suggests that important directions for future work include improving turn models, developing guidelines for their use that are associated with specific research objectives, and demonstrating the robustness of the patterns identified across multiple turn models.

Over the course of this project, we developed and tested a wide array of algorithms to parse a natural conversation into turns. No single model on its own was able to handle all turn-related edge cases that occurred across the entire corpus. Rather, some models appeared to capture certain psychologically intuitive aspects while they missed others. This suggests that either specific research questions will call for the use of specific turn models, or ultimately, further development of turn models will yield something near an automated gold standard that can be reliably deployed across natural conversation to answer a wide variety of research questions.

Table 1. Basic comparison of Audiophile and Cliffhanger turn models. Speakers' mean and median turn durations were four to five times greater for the cliffhanger turn model compared to those of Audiophile. This suggests more broadly that analytic decisions about transcript segmentation will play a key role in the empirical investigation of speaking duration, an understudied topic.

Model	Turn duration		Number of words		Average turns
	Mean	Median	Mean	Median	
Audiophile	2.22	0.92	6.40	2	440.70
Cliffhanger	8.52	5.81	17.81	9	159.41

Backchannels

Backchannels, the short words and utterances that listeners use to respond to speakers without taking the floor (e.g., "yeah," "mhm," and "exactly"), represent a third basic feature of the turn-taking system.

As illustrated in Table 1, the hypergranular Audiophile approach to turn segmentation has limited utility for many research questions. For example, imagine that speaker A tells a 2-min story, during which speaker B nods along and contributes simple backchannels such as "yeah" and "mhm" to demonstrate that they are

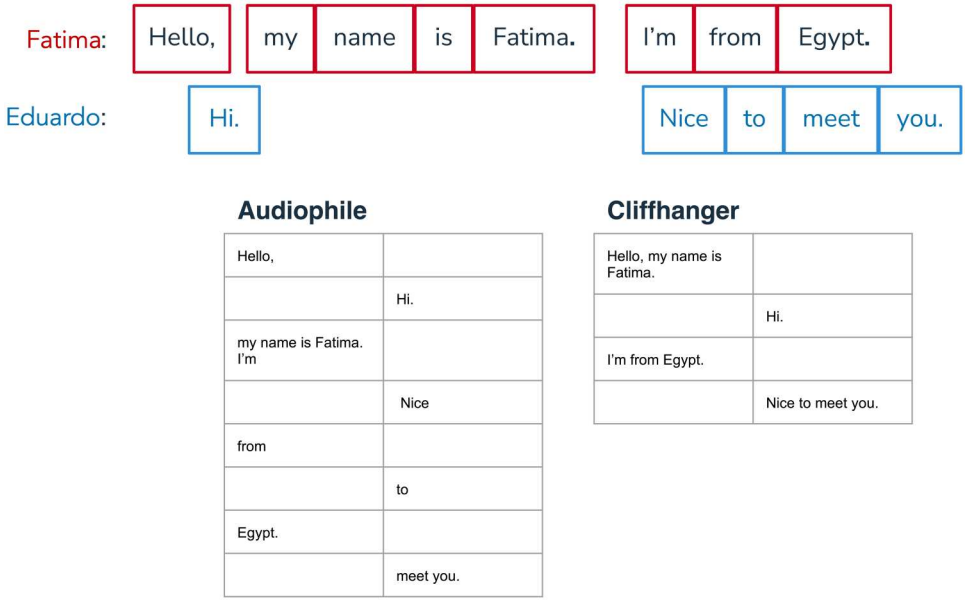

Fig. 3. A depiction of turn segmentation by the Audiophile and Cliffhanger turn models. The baseline Audiophile model treated each interjection as initiated a new turn and, thus, disrupted the flow of Fatima's self-introduction (red). In contrast, our improved Cliffhanger model organizes the same information into a more intuitive format in which Fatima and Eduardo (blue) alternate pleasantries.

paying attention and engaged. In a maximally granular formulation of turn exchange, such as Audiophile, each "yeah" is recorded as a distinct partner turn. This may be inappropriate for certain research questions because it permits the interruption of a continuous narrative based on small, routine interjections. To accommodate such research questions, we developed Backbiter, a turn model that allows the listener's encouraging "yeahs" to simply exist in parallel, rather than interrupting the storyteller's 2-min monolog. Such a model quantifies backchannels as informative and important, but nonetheless peripheral, annotations to primary speaking turns.

Backbiter creates two transcript entries for each speaking turn: first, the words that the turn's active speaker utters and, second, the backchannel phrases that the turn's listener utters, if any. Backbiter uses three basic rules to identify and reclassify utterances as backchannel turns: (i) A backchannel turn must be three words or fewer, (ii) it must contain >50% backchannel words (e.g., "yeah," mhm," and "exactly"), and last, (iii) it must not begin with a prohibited word, such as "I'm..." (see the Supplementary Materials for a complete list of backchannel words and nonbackchannel beginnings.) Using these rules, the Backbiter algorithm moves utterances deemed as backchannel turns from the main turn registry into the turn's accompanying backchannel registry. As shown in Fig. 4, Backbiter altered the transcript considerably by identifying and removing backchannel turns successfully.

A single "mhm" may appear eminently forgettable, but when well-timed and particularly in the aggregate, backchannels become essential conduits of understanding and affiliation (*47*). In our corpus, we observed that listeners deployed backchannels universally: 33.7% of speaker turns elicited at least one listener backchannel. This figure rose to 65.5% of speaker turns that were five words or longer; we estimate a rate of approximately 1000 backchannel words per hour of spoken conversation, which is consistent with previous observations (*48*). Notably, among many possible backchannel words, one reigned supreme: The word "yeah" alone accounted for nearly 40% of all backchannels, either in singular form ("yeah"), double ("yeah, yeah"), triple ("yeah, yeah, yeah"), and more ("yeah, yeah...yeah, yeah, yeah"). (See Fig. 5 for distribution of backchannel frequencies.)

Conversation researchers have identified various ways in which not all backchannels are created equal. For example, one distinction is between "generic" and "specific" backchannels. Generic backchannels, sometimes referred to as "continuers," display understanding and function as a signal to the current speaker to continue talking. On the other hand, specific backchannels, sometimes referred to as "assessments", respond to the content of the current speaker's turn and usually display a degree of affiliative alignment, often by mirroring an emotion of some sort, such as saying "yuck" at the climax of a disgusting story (*49*, *50*). Hence, while the generic "yeah" dominates our corpus in sheer frequency, the use of specific backchannels, such as "wow," may actually be more important in establishing a social connection or in guiding a narrative's direction. A second example is that while "mhm" may function to signal continued understanding, or "passive receptivity," as it has been called, researchers have also suggested that backchannels can signal "incipient speakership" or a listener's readiness to take the floor (*51*). These examples barely scratch the surface of all the various functions of backchannels, and a large-scale empirical investigation is lacking [for a review of backchannel functions, see (*48*)]. Our corpus represents a resource that facilitates this line of research.

Note that when we consider all of these different functions of backchannels, it is clear that simple continuers (e.g., "mhm") may be considered as a low-level feature of conversation. However, given that backchannels also serve more socially complex functions—fitting into a broader class of multimodal signals that involve language, gesture, and intonation—they may warrant

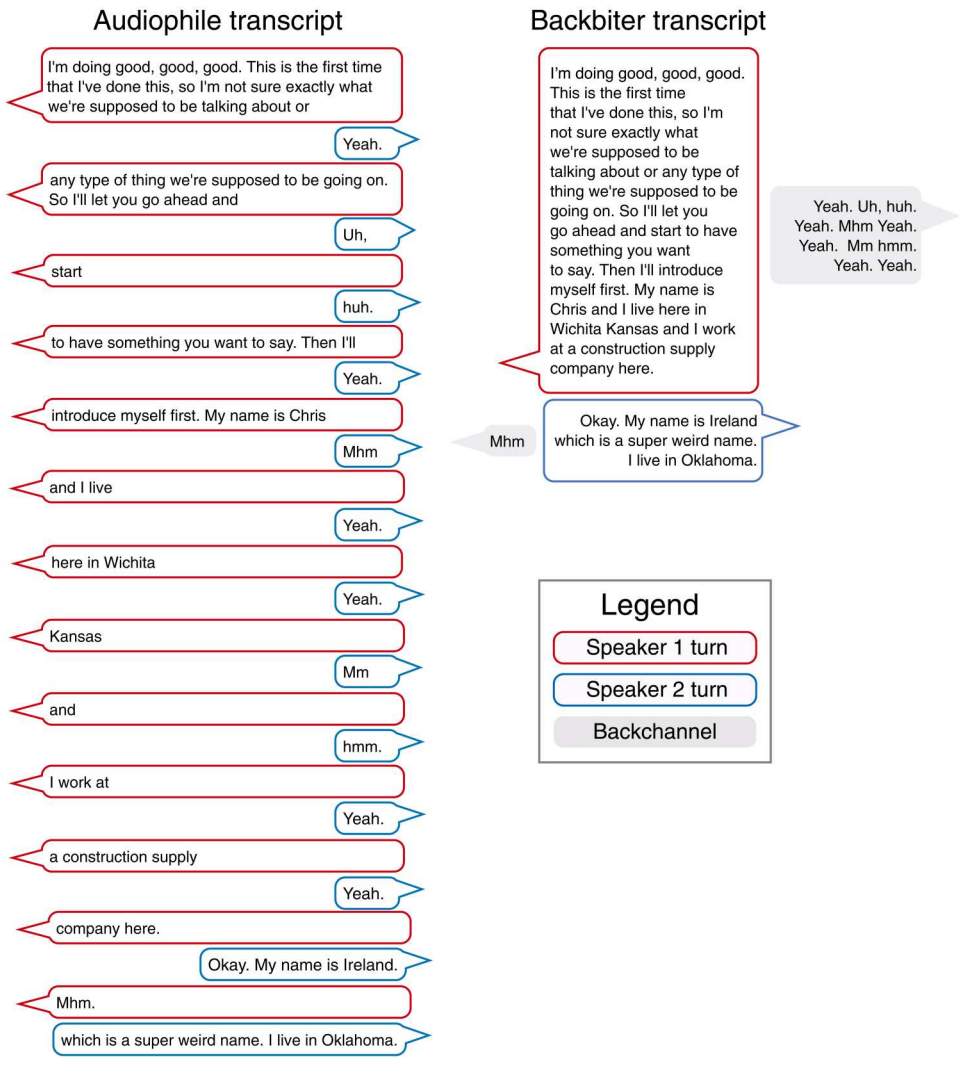

Fig. 4. Example transcripts from Audiophile and Backbiter turn models. Audiophile treats each backchannel as initiating a new turn that disrupts the flow of speaker 1's self-introduction (red). In contrast, our Backbiter turn model organizes the same information and presents it in a more intuitive format in which speaker 1 offers a single introductory turn (while speaker 2 is backchanneling). Speaker 1 then concludes their turn and yields the floor to speaker 2 (blue), at which point speaker 2 takes their first turn and also provides a self-introduction (while speaker 1 occupies the backchannel).

more focused attention as mid-level features in their own right and may require more interpretation, layers of inference, and future research to improve their detection and analysis.

Summary—Turn-taking system

We used the corpus to explore three basic features of the turn-taking system: turn exchange; turn duration, and backchannel feedback. In doing so, we replicated a key turn exchange finding and extended it to the domain of video chat. We also showed that studying turn duration at scale will require developing novel turn segmentation algorithms. Last, we explored backchannels—listeners' widely used signals of understanding and affiliation—and demonstrated the way their automated extraction will likely facilitate further analysis and serve as a starting point for more sophisticated detection algorithms.

Going forward, researchers interested in the causes, correlates, and consequences of conversational turns will benefit from having explicitly defined turn-taking algorithms. Notably, this shift toward the use of automated algorithms in conversation research offers an exciting benefit: By developing a common and explicit language to encode conversational turns, researchers can not only customize turn models to address specific questions but may also share these models with one another to help replicate and extend research, allowing for substantial increases in efficiency. In what follows, we explore high-level features of the corpus and focus on the association between conversation and well-being.

The primary social functions of conversation: Conversation and well-being

The turn-taking system is a fundamental feature of conversation. However, why do humans take such pains to closely coordinate their turns?

Coordination at the turn level ultimately serves cognitive coordination, a state of "intersubjectivity," or shared minds, which lies at the heart of the conversation system. The turn-by-turn

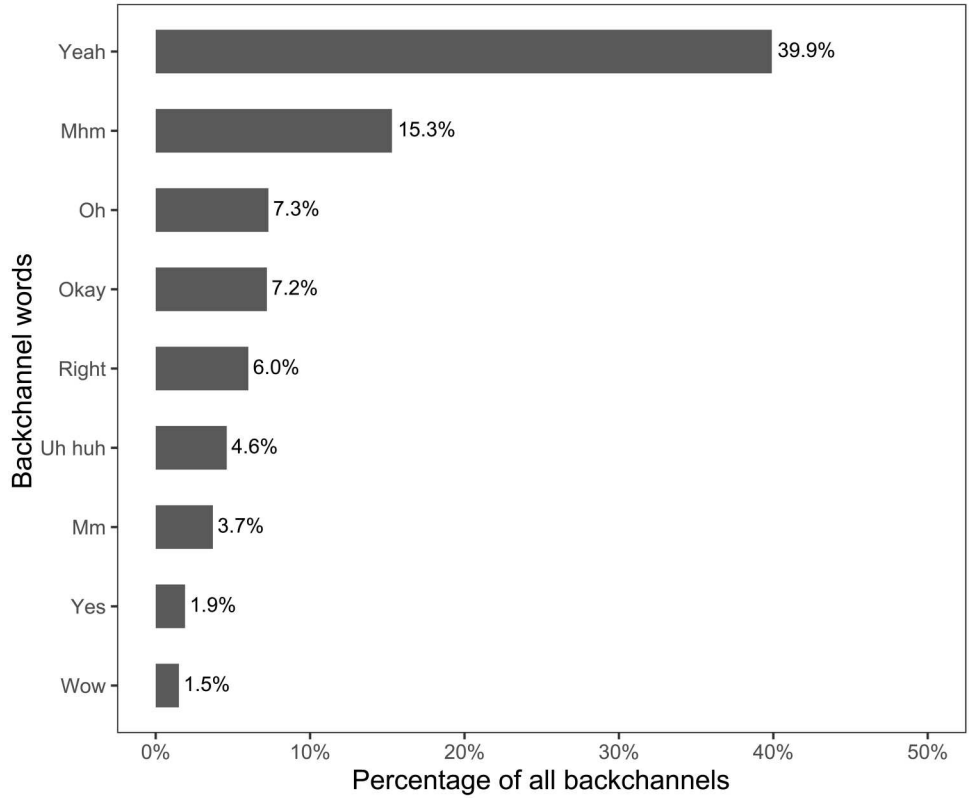

Fig. 5. The frequency of backchannel words across the corpus Backchannel words are a foundational element of conversation that occur at an approximate rate of 1000 per hour of speech; listeners deployed them in nearly two-thirds of speaker turns that were five words or longer. "Generic" continuers, such as "uh huh," may function to signal to speakers that they should keep talking. In contrast, "specific" backchannel words, such as "wow," may convey context-specific responses such as mirroring a speaker's emotion while a story is told. This distribution of backchannels reflects English spoken in the United States in 2020.

coordination of mutual understanding allows people to pursue a range of goals, from persuading, to negotiating, to mating, and so forth. Historically, conversational goals have been divided—not only somewhat simplistically but also somewhat usefully—into "informational" and "relational" goals (*52*). This distinction highlights the different norms that govern such goals, the direct feedback that accompanies information transmission—e.g., "Hold on, what do you mean?" captured, for example, by Grice's (*53*) famous maxims—versus the indirect feedback that often accompanies relational goals, e.g., "I wonder if my new conversation partner really likes me or if they are just being nice?" that certain theories of politeness have described (*54*).

In the CANDOR corpus, people are pursuing informational goals, but the turn-by-turn exchange of information in naturally occurring conversation is more difficult to study at scale. Many other corpora, particularly those that are task-based, offer opportunities for researchers to examine the way people coordinate knowledge. Nevertheless, a comparative advantage of this corpus is that it provides a clear opportunity to view the way people initiate new relationships, form impressions of their conversation partners, and connect with others socially. Accordingly, we examine a question related to relational goals in this section. Note that, as more datasets become available, larger-scale exploration of the interplay between informational and relational goals will become possible, particularly through the lens of certain processes that are central to these goals, such as "repair," which is one of conversation's mechanisms to correct informational misunderstandings [for a review, see (*55*)], and processes that establish social connection, such as "linguistic alignment," the tendency for conversation partners to mirror each other's linguistic expressions [see (*56*, *57*) for an example of the way informational and relational goals can be studied at scale).

In summary, conversation fulfills many goals, including the exchange of information, but one primary objective is the formation and maintenance of social relationships (*12*, *13*). One core finding from a past work is that social interaction, mediated largely through conversation, plays a critical role in people's physical and mental health [e.g., (*15*, *58*–*60*)]. Our corpus offers the opportunity to examine this idea in the earliest stages of developing relationships—people meeting for the first time.

Pre-post positive affect

To explore conversation's hedonic benefits, we had participants report their general mood immediately before and immediately after their conversation by responding to "To what extent do you feel positive feelings (e.g., good, pleasant, happy) or negative feelings (e.g., bad, unpleasant, unhappy) right now?"

A mixed-effects model with random intercepts for participant and conversation revealed that postconversation affect (M = 7.32) was significantly greater than preconversation affect [M = 6.12, b = 1.20, 95% CI = [1.15, 1.25], $t(4085)$ = 44.43, P < 0.001]. This pre-post affective benefit held true across age groups (P < 0.001; see Fig. 6). Thus, at least for individuals who voluntarily opted to engage in a conversation with a stranger, the act enhanced their

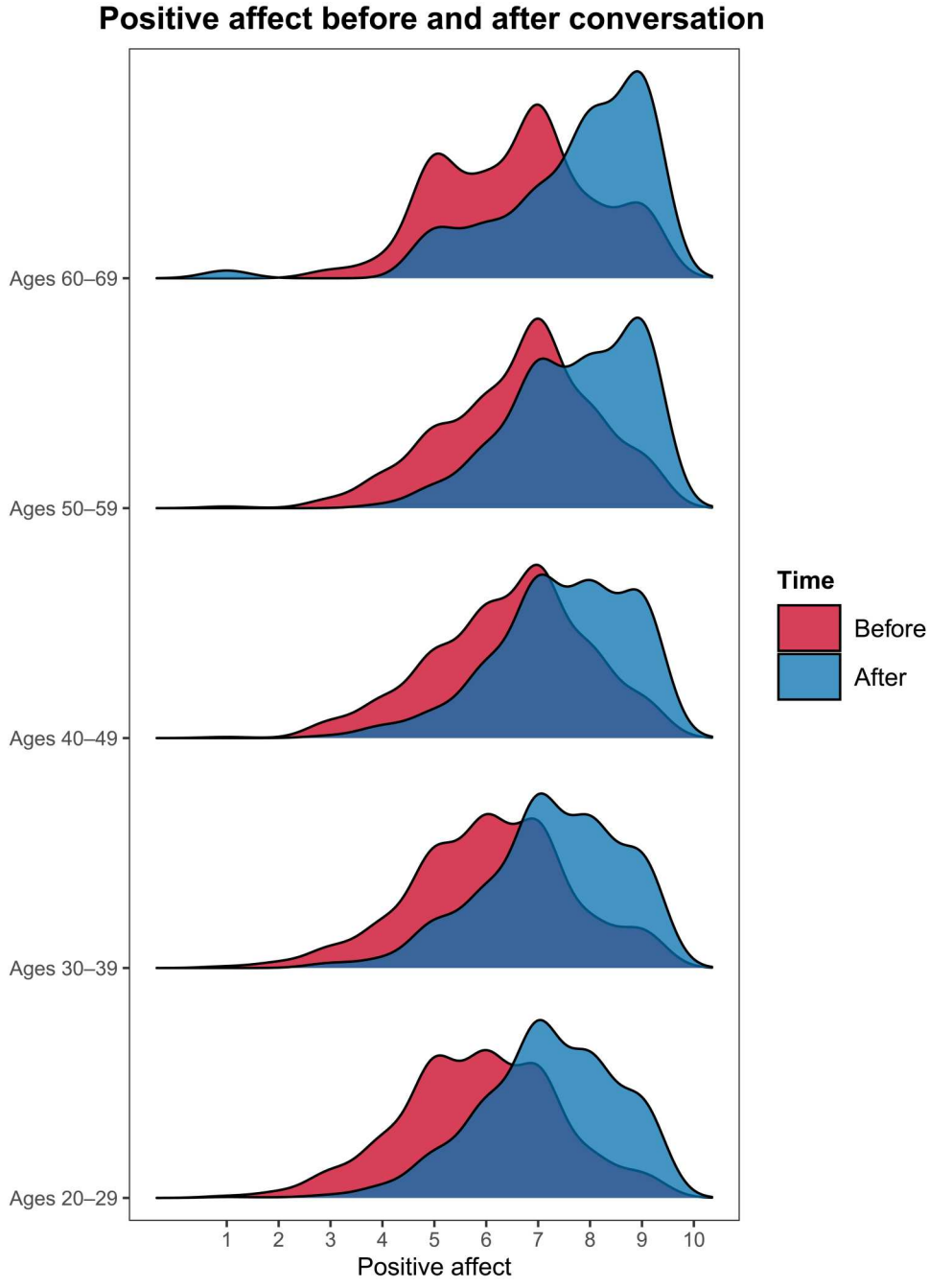

Fig. 6. Positive affect is significantly greater after than before a conversation. Each row of density plots corresponds to an age group. Respondents were asked to report their mood immediately before (red) and after (blue) their conversation. Conversation's effect on people's mood was positive, significant, and of considerable magnitude.

positive affect significantly and consistently across much of the adult life span. Overall, the corpus provides a powerful estimate of conversation's effect on well-being.

Note that, while this effect is consistent with a large existing literature on the positive effect of social interaction on well-being, we cannot entirely rule out that people's repeated answering of the same question about their well-being accounted for some part of the overall effect. Interested readers can refer to the Supplementary Materials for additional analyses that provide evidence against such a demand effect, further bolstering the conclusion that the intervention itself (i.e., people's conversation) is the primary driver of the observed increase in well-being.

While unscripted conversations with a stranger appeared to noticeably improve people's mood, at least for those who opted to engage in them, recent research has revealed a number of mistaken beliefs that people have about such early conversations [for a review,

see (*61*)]. For example, when people were asked about future conversations with a stranger, such as the conversations studied here, they consistently (and incorrectly) anticipated that their conversations would be less interesting, enjoyable, and valuable than they actually were (*62*, *63*). Moreover, after such conversations ended, people were overly pessimistic about what their conversation partners thought of them, including not realizing how much they were liked or the extent to which their conversation partner enjoyed their company (*64*–*66*). Hopefully, large-scale public datasets, such as this corpus, can shed new light on the intriguing social cognition that arises when people use conversation to pursue the fundamental goal of forming and maintaining social bonds and specifically how people's high-level cognition that arises after their conversations end might be related to the dynamics of what was going on at the turn level during their conversations. Moreover, with some additional feature engineering, the interplay between people's broader informational and relational goals may be explored.

Conversation analysis across multiple levels

The previous results demonstrated the range of the corpus and covered conversation's low-level turn-taking features and higher-level outcomes, such as people's overall enjoyment and well-being, respectively. The corpus's breadth uniquely allows analyses that bridge multiple levels, presented here and in the following section.

A limited body of work has probed the associations across different "levels" of conversational analysis, such as influential research on the way listener backchannels determine the quality and trajectory of a speaker's story (*49*). However, such work is the exception rather than the norm, and countless empirical questions remain unanswered. This is attributable in part to the difficulty of collecting conversational data, the inadequate computational methods to quantify features at scale, and the lack of cross-disciplinary collaboration. Below, we demonstrate the way the rapid elimination of these obstacles has now placed this type of research within closer reach.

Turn exchange and conversational enjoyment

In an attempt to associate the lower-level mechanics of conversation with higher-level psychological outcomes, recent research has begun to investigate the way the interval between people's turns may function as an honest signal of whether two conversation partners "click" and enjoy each other's company (*67*). Here, we replicate and extend these previous results.

Following the same measurement approach used in the "Turn Exchange" section (*35*), we defined a person's mean turn interval based on the durations between their conversation partner's turn endings and their own turn beginnings. Participants reported how much they enjoyed the conversation on a nine-point Likert scale (end points: "1 - Not at All" and "9 - Extremely"). We then regressed how much people's partners enjoyed the conversation on people's mean turn interval. This analysis revealed that as the mean turn interval decreased, the partners' enjoyment increased [$b = -0.73$, 95% CI = [−0.92, −0.53], $t(3255) = -7.17$, $P < 0.001$]. The same analysis using the median turn interval yielded a similar result [$b = -0.69$, 95% CI = [−0.91, −0.47], $t(3255) = -6.07$, $P < 0.001$]. Thus, the faster people responded when taking the floor, the more their partners enjoyed the conversation. However, per our earlier discussion (see Turn Exchange section), turn intervals come in two different types—gaps and overlaps—which are conceptually, and likely psychologically, distinct. We can use this information to expand upon the relationship between turn intervals and conversational enjoyment.

By indicating in our model whether a person's mean turn interval was a gap or an overlap, we were able to examine whether the relation between turn interval and enjoyment differs on the part of people who are, on average, "gappers" versus "overlappers." We found that a significant turn interval × interval type (i.e., gap versus overlap) interaction moderated the relationship between turn interval and partner enjoyment [$b = 0.74$, 95% CI = [0.13, 1.36], $t(3253) = 2.38$, $P = 0.02$]. Postestimation analyses revealed that, for people whose mean interval was an overlap, there was no discernible relation between the duration of their overlaps and their partners' enjoyment [$b = -0.26$, 95% CI = [−0.71, 0.20], $t(3253) = 1.11$, $P = 0.27$]. On the other hand, for those whose mean interval was a gap, there was a significant negative relation between gap duration and partner enjoyment [$b = -1.00$, 95% CI = [−1.41, −0.59], $t(3253) = -4.76$, $P < 0.001$].

Rather than averaging people's turn intervals for each conversation, we also modeled all of their turn intervals across each conversation. To account for the data's structure, we used a linear mixed effects model, with turn interval and interval type (e.g., gap or overlap) as fixed effects and a random intercept and slope for participant ID. This analysis also revealed the same significant interval × interval type interaction [$b = 0.02$, 95% CI = [−0.03, 0.004], $t(331{,}854) = -2.58$, $P < 0.01$]. This finding reinforces the main conclusion, although future work may need to improve the modeling of within-person and within-conversation variance.

In short, as shown in Fig. 7, people's turn intervals were related to how much their partners enjoyed the conversation. However, it was not simply the case that the more rapidly one responded, the more one's partner enjoyed the conversation; the relation between longer gaps and lower enjoyment appeared to determine the effect.

The preceding analysis is just one example of the countless questions that remain about the influence of low-level structural factors on people's high-level impressions of their conversations. One direction for future work is evaluating the stability of people's turn intervals, or the extent to which people's pattern of turn-taking functions as a trait over time, by examining speakers who had multiple conversations. Future research might also manipulate not only intervals but also turn duration and other conversation mechanics experimentally to assess their causal effects on enjoyment and other high-level judgments and impressions. Last, while we used Heldner and Endlund's (*35*) turn model for consistency with the section on the turn exchange, an important future direction would be to consider its robustness compared to alternative segmentation algorithms.

Machine learning and what distinguishes a good conversationalist

Now, we move from the low-level structure of turns and the high-level postconversation outcomes examined previously to the mid-level content of conversation. The breadth, scale, and detail of our corpus offers an unprecedentedly rich view of the way conversation unfolds—moment to moment and turn by turn, through text, audio, and video modalities—across more than 7 million words and 50,000 min of recordings. Here, we introduce and analyze mid-level features such as turns' semantic content, dynamic vocal prosody, and facial expressions. Historically, these factors have

Fig. 7. Turn exchange is related nonlinearly to partner enjoyment. The *x* axis indicates an individual's mean interval between the end of their conversation partner's turn and the beginning of their turn. Positive intervals indicate gaps between turns, and negative intervals indicate overlaps in speech near turn boundaries. We found that longer positive intervals (gaps) between turns were negatively associated with partner enjoyment, but we found no such relation between enjoyment and negative intervals (overlaps). This underscores the importance of connecting lower-level features of conversation with higher-level features, and the interdisciplinary understanding required to do so.

been difficult to analyze because of the labor intensiveness of annotation, the subjectivity of their perception, and the high-dimensional nature of textual and audiovisual data. However, recent advances in speech analysis and machine learning have increasingly allowed scholars to investigate these nuanced aspects of social interaction. We explore this rich "middle layer" of interaction and associate it with high-level impressions by exploring an open question in conversation research: What distinguishes a good conversationalist?

We begin by using a suite of open-source, audio-processing, and computer vision models to extract detailed, high-frequency audiovisual information—features such as head pose, speech spectrum, and so forth—for each moment in our nearly 850-hour corpus. These fine-grained measurements were then transformed into turn-level features. Overall, our analysis characterized the linguistic, auditory, and visual content of 557,864 conversational turns (using the Backbiter turn model) along 19 dimensions (see the Supplementary Materials for full details).

Our analyses revealed substantial differences in semantic novelty, vocal dynamism, and facial engagement, in the way good and bad conversationalists—according to their partner's evaluation—communicate. We also identified numerous avenues where additional human annotation, refinement of computational models, or application of domain transfer techniques may help advance the study of conversation.

Characteristics of good and bad conversationalists. After a conversation concluded, each participant was asked to rate their partner as follows: "Imagine you were to rank the last 100 people you had a conversation with according to how good of a conversationalist they are. '0' is the least good conversationalist you've talked to. '50' is right in the middle. '100' is the best conversationalist. Where would you rank the person that you just talked to on this scale?" In general, participants reported that their conversation partners were above-average conversationalists (mean = 73.0, SD = 20.1). For simplicity, we defined "good" and "bad" as the top and bottom quartiles of the partner-rated conversationalist scores. The main text results focus on the contrast between these groups; complete results on middle quartiles (25th to 50th and 50th to 75th percentiles) are provided in the Supplementary Materials.

Here, we present the results for six turn-level features that illustrate the breadth of the analysis: (i) speech rate, i.e., words per second; (ii) the semantic novelty of a speaker's current turn compared to their partner's previous turn; (iii) loudness; (iv) vocal intensity; (v) nodding "yes" and shaking "no" while listening; and (vi) happy facial expressions while listening. Each of these features is depicted in one panel of Fig. 8, in which the top, middle, and bottom rows represent the text, audio, and video modalities, respectively. The left column—speech rate, loudness, and head movement—are features that can, at least in principle, be observed directly. By

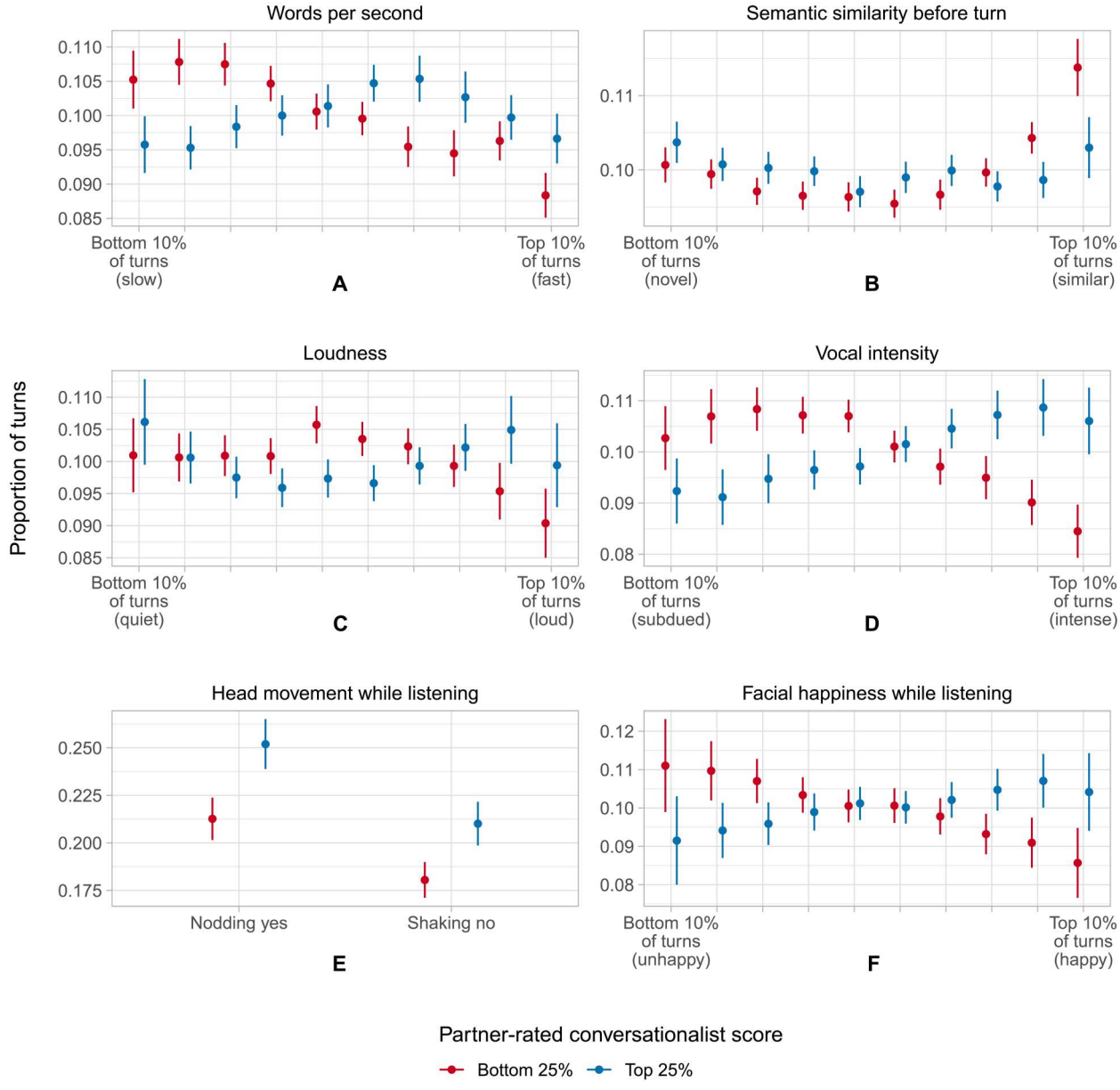

Fig. 8. Behavior patterns of good and bad conversationalists. (**A** to **F**) The behavioral patterns of good conversationalists (top 25% of partner-rated conversationalist score; depicted in blue) and bad conversationalists (bottom 25%; depicted in red) are depicted. Horizontal axes denote turn-level feature deciles. The *y* axis indicates the mean proportion of turns in a category for a good or bad conversationalist. Error bars represent 95% confidence intervals. Top, middle, and bottom rows correspond to text, audio, and visual modalities, respectively; left and right columns include features that can be observed directly and those that require an additional layer of machine learning to estimate.

contrast, the features in the right column—semantic novelty, vocal intensity, and facial expressions of happiness—are more complex, subjective mid-level concepts that require an additional layer of machine learning inference to proxy.

People's speech and behavior varied considerably over the course of a half-hour conversation, and these complex patterns were difficult to capture with simple linear analysis. Instead, we examined whether good and bad conversationalists had different distributions across features. To facilitate presentation, we represent feature distributions as frequency plots binned by deciles. For example, Fig. 8A shows the frequency, indicated by the proportion of conversational turns, at which good and bad conversationalists spoke very slowly (i.e., in the bottom 10% of speaking speeds), and so forth, up to the fastest speech rate (i.e., in the top 10% of speaking speeds). To quantify the trends' general direction, we also report differences in means (see the Supplementary Materials for full details of the statistical procedure).

This approach allowed us to examine the way good and bad conversationalists varied in the proportion of turns they took that demonstrated more or less of any given feature. In what follows, our analysis, visualization, and interpretation of the features follow this analytic approach across six selected features and reveal distinct

patterns in the way good and bad conversationalists engage with their partners.

Speech rate

We begin with speech rate, a low-level feature that is both straightforward to calculate and is associated with traits that are conceivably related to one's ability as a conversationalist, such as competence (*68*), persuasiveness (*69*), and intelligence (*70*). For each turn, we computed speech rate by dividing the number of words spoken in each turn by that turn's duration in seconds, which yielded a rate of spoken words per second (WPS). For analysis, WPS were binned into deciles per the procedure outlined above. Figure 8A shows that good conversationalists spent more of their turns speaking quickly (i.e., in the upper five deciles). In contrast, bad conversationalists spent a greater proportion of turns speaking slowly (i.e., in the lower four deciles).

After adjusting for multiple comparisons, the speech rate distributions of good and bad conversationalists differed significantly; the null of equal distributions was rejected at adjusted *P* value (P_{adj}) < 0.001. In summary, Fig. 8A shows that good conversationalists spent more time speaking quickly, while bad conversationalists spent more time speaking slowly. On average, good conversationalists spoke at a rate of 0.1 WPS faster than bad conversationalists (95% CI [0.06, 0.14]). For comparison, the mean speech rate across our corpus was 3.3 WPS, indicating a 3% increase in speed.

Semantic similarity

While speech rate is a straightforward property of a conversational turn (assuming a time-stamped transcript from which to begin), the semantic exchange between two speakers is a more nuanced and psychologically complex aspect of conversation. Using machine learning methods for dimension reduction, we computed a measure of semantic content based on text embeddings. Text embedding techniques numerically represent sentences or documents as vectors, based on co-occurrence patterns in a large training corpus; the resulting representations are widely used to measure the semantic distance between words, sentences, and documents (*71*). This provided a proxy of each speaker's novel "contribution" in any given turn, relative to the previous turn in a conversation.

Consider the following excerpt from a conversation in our corpus between two speakers who we will refer to as D and K. Following a relatively boring turn on D's part that threatens to stall the dialog, K shifts skillfully to a fresh line of inquiry.

D: [Talking blandly about the weather] "It's the same, I think it's the same here, about 32 degrees."

K: [High semantic novelty response] "Yeah, exactly. Okay. It's going to appear like a totally random question, but being a Wisconsinite, how frequently do you attend fish fries?"

In this turn couplet, K responded to her partner's floundering statement about the weather with a novel question about attending Wisconsinite fish fries, which reset the conversation's momentum.

To study this computationally, we generated turn-level text embeddings from corpus transcripts with MPNet, a pretrained language model that achieves top performance currently on a variety of linguistic tasks [(*72*); via the Sentence-Transformers Python module, from (*73*)]. We then used the cosine similarity between the embedding vectors of (i) the current turn and (ii) the turn immediately before and obtained a proxy for the degree of semantic novelty injected into the conversation.

Figure 8B shows that good and bad conversationalists differ significantly in their turns' novel semantic content (null hypothesis of equal distributions rejected at $P_{adj} = 0.001$). The results are robust to an alternative Euclidean distance metric, as well as to the widely used RoBERTa embedding model (*74*); see the Supplementary Materials for details. However, as Fig. 8B makes clear, it is not the case that good conversationalists add more novelty to their turns across the board; rather, they use a mix of semantically novel and semantically similar turns. Despite some caveats (see section S3.4), we report these results on textual novelty because of their robustness across multiple specifications and unsupervised machine learning techniques' apparent ability to capture nuanced aspects of conversational skill that could not be studied computationally until recently.

Loudness

While the exchange of semantic content plays a central role in conversation, a vast literature has also established that paralinguistic cues, such as vocal tone, are similarly important. The low-level acoustic characteristics of speech can be quantified in many ways; we focus here on loudness (as measured in decibels or log-scaled vocal energy). Note that "loudness" (the perceptual strength of sound) is what most people refer to colloquially as "volume." Technically, volume is the auditory sensation that reflects the size of sound, from small to large, or, less formally, the "bigness," "spread," or "space-fillingness" of sound (*75*, *76*). Louder speech often attracts a listener's attention effectively, although its overuse can backfire (*77*). Anecdotally, speakers who vary their volume for emphasis may be considered more dynamic, while those with a uniformly loud or quiet voice may be perceived to be monotonous. To examine these patterns in our corpus, we computed per-turn loudness values and used those to compute interturn variation in loudness (see the Supplementary Materials for additional analyses of intraturn modulation of loudness, as well as an analysis of pitch).

As Fig. 8C shows, our analysis revealed that good and bad conversationalists differed significantly in the distributions of the loudness of their turns ($P_{adj} = 0.03$). Differences in loudness distributions persisted when the first and last seconds of a turn were clipped or turn duration was adjusted for linearly (see the Supplementary Materials for details). In additional analyses that were disaggregated by speakers' gender, we found that male speakers primarily were responsible for these results. Moreover, consistent with intuition, loudness patterns were highly nonlinear: We found no significant difference in the mean turn loudness between good and bad conversationalists. Rather, bad conversationalists spend more time taking turns that are of medium loudness, while good conversationalists spend more time taking turns with either lower or higher average loudness—and perhaps thereby match the dialog's needs more adeptly? In short, using one's voice to occupy a range of loudness values appears to be associated with conversational skill, but additional research is required.

Vocal intensity

Although loudness is appealing as an important acoustic feature that can be measured transparently, humans often use more complex combinations of vocal characteristics, including roughness, sibilance, or the contrast between lower and higher frequencies, to convey information such as emotion (*78*). One such acoustic amalgam is emotional "intensity" (sometimes referred to as "activation"), a basic property of emotion and momentary affect (*79*, *80*). As with other concepts perceived subjectively, the precise definition of intensity is debated; previous work has related it to changes in one's body, how long the emotion lingers, the degree to which it

motivates action, and whether it changes one's long-term beliefs (*81*). For our purposes, we used a measure of acoustic intensity to test the hypothesis that spoken intensity differs among good and bad conversationalists.

To do so, we used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (*82*) to train a vocal intensity classifier and then applied this model to assign intensity scores for each speaking turn in our corpus. The RAVDESS dataset consists of recordings of trained actors who were prompted to read simple statements with either "normal" or "high" emotional intensity; we treated this intensity label for each recording (normal or high) as the response variable in training our classifier (a logistic regression model). Model predictors consisted of summary statistics for each RAVDESS recording across a range of common prosodic features: mean, maximum, and SD for fundamental frequency (F0) and volume (log energy), as well as voiced and unvoiced duration (see the Supplementary Materials). We used our resulting trained classifier to predict vocal intensity for every 1-s interval in the corpus and then averaged these values within turns to obtain a single intensity score per turn.

As Fig. 8D reveals, people rated as good conversationalists spoke with greater intensity than bad conversationalists; the null hypothesis of equal distributions was rejected at $P_{adj} < 0.001$. These results were fairly linear (difference in means of 1.3% points in the predicted probability of high intensity, 95% CI [0.01, 0.02]) and were also robust to an alternative specification that adjusted for turn duration. In analyses disaggregated by the speaker's gender, differences persisted among female speakers, but became not significant among male speakers (see the Supplementary Materials). Whether listeners consider vocal intensity enlivening, emotional, or empathetic remains an open question.

Similar to other model-based predictions for subjective mid-level features, these results are intended simply as a starting point for future work. First, it worth emphasizing that any effect related to complex features such as vocal intensity will likely exhibit considerable variance that has yet to be explored. For example, consider how different personality traits may moderate the relationship between vocal intensity and good and bad conversationalists (see section S3.5.). Second, we used a relatively simple model for exposition and computational efficiency, which leaves room for considerable improvement through additional feature engineering as well as the use of domain transfer techniques and more sophisticated classifiers. To allow continued refinement, we provide raw audio files for all conversations in the corpus, in addition to the tabular records of the extracted features that we selected for analysis. We also provide our vocal intensity estimates for others to replicate. Last, we emphasize that concepts such as intensity are not merely unimodal. In addition to its manifestation in voice, facial expression, word choice, and so forth also communicate intensity, which highlight the need for continued research on the multimodal measurement of emotion expressed in conversation.

Head movement

In the cultural context of this American corpus, a commonplace nonverbal cue of assent or agreement is the up-and-down head nod. Similarly, shaking one's head from side to side usually signals negation or disagreement. To capture head movement patterns in the corpus recordings, we developed an algorithmic "nod detector." Using facial recognition software in the Dlib C++ library (*83*), we computed a set of facial landmarks to identify the position of any human face detected on screen. From there, we developed a rule-based scheme to evaluate whether, over a 2-s period, (i) at least 10% of a participant's face (ii) crossed its beginning position at least twice. When this occurred along the vertical axis, we recorded a "nod." If it occurred along the horizontal axis, then it was recorded as a "shake." Binary summary features were computed separately for nods and shakes to indicate the presence or absence of nodding and shaking at any point in the turn.

The results showed that good conversationalists differed in both of these common nonverbal listening behaviors. Figure 8E depicts the rate of head nodding and head shaking among both groups. We found that good conversationalists were significantly more engaged not only in their rate of nodding "yes" (4.0% point increase; 95% CI [0.02, 0.06], $P_{adj} < 0.001$) but also in their rate of shaking "no" (3.0% point increase; 95% CI [0.01, 0.05], $P_{adj} = 0.001$). Notably, these results indicate that good conversationalists are not merely cheerful listeners who nod supportively at each new contribution from their partners. Rather, they also make judicious use of nonverbal negations (head shakes) when appropriate. This suggests that good conversationalists' head movements during conversation are best characterized as engagement rather than just simple positivity.

As with our other results, we emphasize the need for continued research on the measurement of nonverbal engagement. Informal testing suggested that our nod detector demonstrated good precision (i.e., a low false positive rate) but had weaker recall (i.e., more false negatives than desired). A more refined facial recognition algorithm may be able to detect more fine-grained head movements, as well as to extract additional head pose information (e.g., whether a listener's head is cocked to the side), and these and other improvements will undoubtedly improve performance on this and similar tasks of visual classification.

Facial happiness expression. Last, we turn to facial expressions, a more nuanced visual cue. Unlike nods, which consist of a simple up-and-down movement, the scope of potential facial expressions is large and can be difficult to interpret. To address this challenge, we used an emotion recognition model pretrained on the AffectNet corpus of facial expression images categorized into eight emotional groups (*84*). For every second in the corpus, we provided speaker and listener images to a convolutional neural network that assigned a probability to each emotional label. Note that the perception of facial emotions is highly subjective, as evidenced by AffectNet's low reported intercoder agreement. This is both a technological and conceptual issue, as people express the same emotion differently, often transition quickly between emotions, and even combine aspects of different emotions in idiosyncratic ways [e.g., (*85*, *86*)]. Moreover, the model was not adapted to our video conversation context, where extreme facial contortion is rare and emotions such as contempt, disgust, fear, sadness, and surprise appear to be virtually undetected. We avoided these issues by examining expressions of facial happiness alone, as happiness and neutrality were, by far, the most common expressions detected in our corpus. In what follows, we present results on expressions of facial happiness while listening (see the Supplementary Materials for an additional analysis of happiness while speaking).

Figure 8F demonstrates that good conversationalists exhibited significantly more (difference in mean predicted probability 3.5% points; 95% CI [0.02, 0.06]) facial happiness expressions while listening, compared to bad conversationalists (null of equal distributions rejected at level $P_{adj} = 0.05$). Male speakers were largely

responsible for these results, and they became statistically not significant in an analysis limited to female listeners (see the Supplementary Materials). As with the results on head movement, this finding highlights the role of engaged listening in differentiating good and bad conversationalists.

Across our corpus, we found that good conversationalists were characterized by a number of directly measurable objective behaviors: They spoke more rapidly, showed greater variation in loudness across their speaking turns, and engaged in active listening through nonverbal cues (head nods and shakes). We also identified a number of related patterns among more nuanced and psychologically complex mid-level behaviors that required trained algorithms to detect. Good conversationalists injected more semantically novel content into their turns, exhibited greater vocal intensity while speaking, and exhibited more expressions of facial happiness while listening. Together, these findings demonstrated our corpus's considerable potential to explore conversation in innovative ways, particularly across levels of analysis, which we feel will be an increasingly important type of conversation research.

A qualitative glance at the corpus—Topical, relational, and demographic diversity

While our report focuses primarily on empirical patterns, the corpus also offers a unique lens into American discourse in 2020. Consider that our corpus consists of conversations collected during a hotly contested presidential election and a global pandemic. This makes the dataset a social repository of one of the most unusual stretches of time in recent memory.

As shown in Fig. 9, using simple string matching, one can clearly see the increased discourse about the election (keywords: "election," "biden," "trump," "republican," and "democrat"), as well as the even more marked rise of coronavirus disease 2019 (COVID-19) as a topic of conversation as the pandemic gripped the globe (keywords: "covid," "pandemic," "vaccine," and "mask"). Together with these topics, one can also see the growing summertime focus on law enforcement and police killings (keywords: "taylor," "floyd," and "police"). Last, note the near-universal inclination of people everywhere to want to talk about their family and children (keywords: "my kids," "parents," and "family"). In an effort to explore understudied aspects of conversations and to articulate a larger structural framework, we left questions of topic choice relatively untouched, which we suspect will be a particularly fruitful direction for future research.

Regardless of how much technology is applied, conversational discourse cannot be fully appreciated without actually watching people talk. Over the course of a year, a member of our research team engaged in a close qualitative examination of the corpus's contents and watched every minute of its 850 hours. On the basis of notes taken over the course of this monumental effort, we provide records that describe, among other categories, conversations that were particularly noteworthy, either for their awkwardness or for the astounding degree of rapport that emerged (see the Qualitative Review). Such qualities, which at present are impossible to identify computationally, immediately suggest directions for future research into the characteristics that distinguish these interactions.

Much qualitative research remains to be done. Several research disciplines, including discourse analysis and conversation analysis,

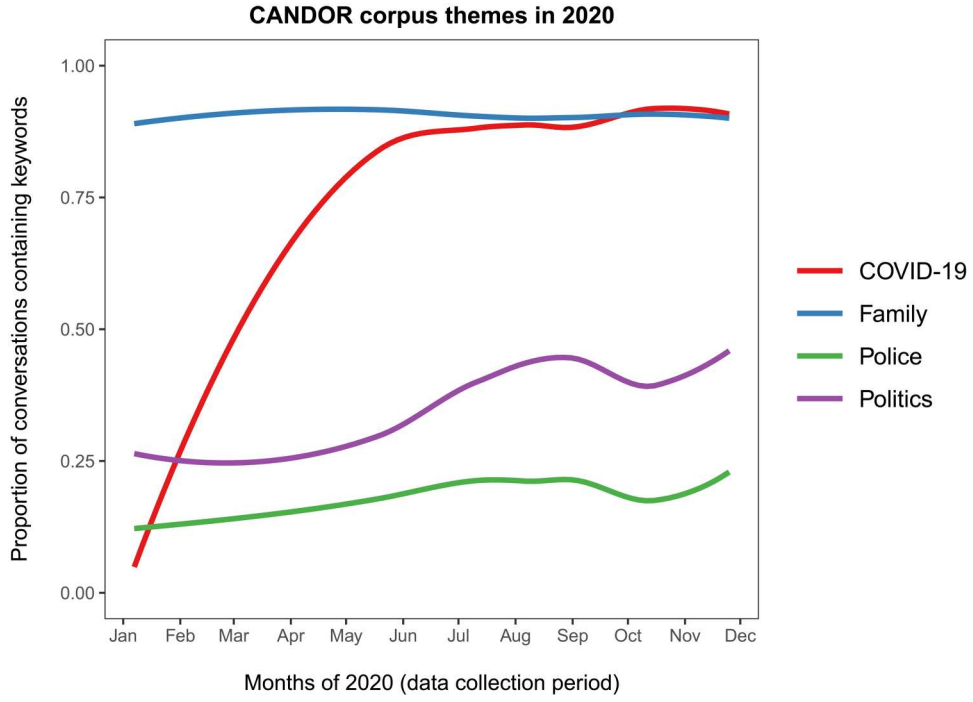

Fig. 9. Topic flow within the CANDOR corpus. The topics people chose to talk about, as measured in CANDOR transcripts by a simple keyword dictionary, reflect the ebb and flow of societal issues in an unusually tumultuous year. COVID-19 (red) surged from unknown to the talk of the nation by mid-2020, matching or even exceeding family-related discussion (blue), a reliable staple of conversation. CANDOR frequencies for the presidential election (purple) and policing (green) highlight the trajectories of these nationally debated issues.

have historically studied conversation using high-quality manual transcription, fine-grained annotation, and close reading of important conversations [for an overview, see (*87*); for an approachable introduction to some key concepts, see (*88*)]. This careful work has yielded a considerable number of fundamental insights into conversation, many of which had a strong influence on the analyses in the section on turn exchange. In making the corpus public, our goal is not only to make further work on this subject feasible but also to benefit from the expertise in these research disciplines to refine turn segmentation algorithms, establish gold-standard measures to train machine learning models, and identify conversational phenomena that may be explored further both empirically and experimentally.

An additional recommendation for future research is to explore the notable diversity in our participant pool, which represents a broad cross section of the United States, particularly in cases where dissimilar participants were paired, resulting in countless intergroup conversations. For example, in one conversation, our qualitative reviewer noted that "… [speakers are] a white man and a black woman, begins off very reticent and one-sided, rapport develops slowly with the help of a kind and talkative partner." In a different conversation, "… [speakers are a] 40-year-old mother of 3 in Louisiana and 20-ish daughter of Polish immigrants located in Chicago. Mother is in recovery and was raised in the foster system. She is illiterate. Discussion about cycles of trauma and privilege and resilience." Watching such conversations, one is reminded that regardless of how advanced our computational techniques become, reducing conversations to tabular rows and columns of data will fail to convey the full richness of their humanity.

The cross-demographic pairings in our corpus also offer a rare opportunity to study the way people navigate a lengthy one-on-one interaction with an unfamiliar conversation partner, who is often someone who appears very different from them. Exploratory analyses revealed the way people's conversational behaviors shifted when they were assigned to partners from different age, gender, racial, educational, and political groups (see the Supplementary Materials for full results). These analyses revealed a number of notable behavioral differences compared to conversations among demographically similar participants.

For example, we found that older speakers, defined as belonging to the upper age tertile, tended to take significantly longer turns when talking to a younger conversation partner compared to when talking to another older speaker (1.5 s, 95% CI [1.1, 1.8], $P_{adj} < 0.001$). This effect was large and constituted an approximately 9% increase in floor time for older participants in age-mismatched conversations compared to age-matched conversations. A second comparison revealed that, among female participants, vocal expressiveness (measured as the SD of vocal pitch) increased by 1.2 Hz when speaking with female partners, compared to a baseline SD of 37.7 Hz when speaking to males (95% CI [0.6, 1.8], $P_{adj} = 0.01$) [see (*89*)]. Last, we observed that white participants used 15% fewer backchannels when paired with Black partners compared to when paired with white partners ($M = 0.01$ fewer instances per second, 95% CI [−0.02, −0.01], $P_{adj} < 0.001$). This is consistent with previous research results that showed that white participants often decrease nonverbal signals in interracial interactions (*90*).

We urge caution in interpreting these results, as apparent intergroup differences should not be attributed solely to the perceived outgroup aspect of a conversation partner's identity. Social identities are complex and often contain components that are difficult to disentangle [see (*91*) for a discussion of causal inferences about bundled identities]. Nonetheless, our corpus does contain a wide diversity of dyadic pairings and, hence, may lend itself to any number of interesting future lines of inquiry with respect to intergroup communication.

DISCUSSION

To guide our exploration of the CANDOR corpus, we divided conversational features into three levels: (i) lower-level objective features of conversation that can be measured directly at high frequency; (ii) mid-level, psychologically rich features that can be inferred indirectly, often because of advances in machine learning; and (iii) higher-level subjective impressions reflected in participants' postconversation survey responses. We first explored these levels in isolation, examining low-level features of the turn-taking system, followed by high-level features, such as people's reported well-being. Then, we used the corpus to draw connections across levels of analysis—an exciting form of conversational research that opens numerous lines of inquiry, many of which will require interdisciplinary collaboration. In doing so, we examined the relation between people's speed of turn exchange and their partners' enjoyment; we also explored the rich middle layer of conversation and found that various mid-level features, including semantic similarity, acoustic intensity, and facial expressions of emotion, were able to distinguish good conversationalists from bad ones.

In the most casual interpretation, these "levels" simply help organize a vast, multifeatured dataset into convenient categories to report analyses that belong to different families of content. However, we propose that this notion of a conversational hierarchy, in addition to its practical utility, may also prove fruitful in generating theoretical insights (please see the Supplementary Materials for an extended discussion).

Practical considerations: Exploring the corpus

While constructing the CANDOR corpus, we encountered a number of interesting challenges, ranging from technical (e.g., aligning video and engineering features) to conceptual (e.g., defining psychologically sensible turns and representing conversation in terms of "levels" of analysis). Throughout this report, we have sought to justify the decisions required to make analytic progress, while reasoning transparently through limitations, alternative approaches, and the potential downstream consequences. We invite readers to examine our choices (and omissions) and, ideally, improve upon them in future research.

Politics and the pandemic

Our corpus consists of conversations collected during a contentious year in America at the onset of a global pandemic. This makes the dataset a fascinating social repository of historical record. At the same time, as a reference for conversation science, it is important to consider the context, particularly when asking certain questions, such as those related to politics, topic choice, feelings of social isolation, questions of universality, and so forth.

Unique sample

Our participants were English speaking and resided in the United States, which represents the living conditions of only a fraction of the world's population (*92*). Our sample also consists of people willing to talk to strangers online, which may be their most

unique feature of all. Ultimately, we will have to wait for other corpora to be released to make progress on important cross-cultural comparisons.

Dyadic interaction

If the science of dyadic conversation is incomplete, then that is nothing compared to the lack of research on group conversation (*93*–*95*). Unfortunately, our dataset does not fill this gap, although it may serve as a starting point for future studies, for example, by establishing robust baseline values for dyadic conversation to contrast with future work on groups.

Talking with strangers

Our dataset consists of conversations between people who had never met before. At the very least, this indicates that certain conversational phenomena will be underrepresented. For example, gossip, which often functions to reinforce people's social bonds, occupies a considerable portion of talk time in everyday conversation (*13*, *96*). Two strangers can certainly still "gossip" in the colloquial sense of talking about celebrities, for example, but the gossip that truly dominates daily conversation relates to mutual acquaintances who are not present. Similarly, our dataset may be less suited to examine a phenomenon such as self-disclosure, which occurs to a greater extent in intimate relationships, such as between close friends and significant others [e.g., (*97*)], although our Qualitative Report documented an unexpected number of conversations that included deep disclosures.

"Getting to know you" conversations

Our participants were not given specific instructions; rather, they were simply told to have a conversation. This has the benefit of producing a corpus of natural conversation that complements existing task-specific datasets or corpora consisting primarily of institutional talk. On the other hand, to investigate certain conversational phenomena, researchers may benefit from more structured forms of talk. For example, researchers interested in the way people cooperate to create common ground and mutual understanding may be served better by conversations that involve completing a joint communication task (*98*).

Video chat

As video chat is becoming a dominant communication medium, much work remains to be conducted on its effects on perceived eye contact, facial expressions, turn-taking, impression formation, and so forth. Questions also remain about which behavioral patterns may be inherent to the digital medium in general, rather than dependent upon specific aspects of the medium, such as internet speed or camera resolution. While many phenomena may ultimately prove to be medium-independent (for example, the turn-taking results were remarkably consistent with face-to-face results), scholars should be cautious in extrapolating these results to other contexts. Nevertheless, as society rethinks human communication, including a move toward more digital communication, remote work arrangements, and so forth, understanding video-mediated conversation is an increasingly important endeavor.

Downtime

At the beginning of each conversation, one person signed on before their partner. During this downtime, extraneous signals sometimes entered the recording: background noise captured as speech, facial expressions, and even people talking to themselves. Hence, when calculating aggregate statistics, we recommend using only the period when both speakers have appeared (approximated reasonably by the beginning of the second transcript turn). Overall, when calculating statistics, such as smiles per minute, caution should be exercised in the choice of one's denominator.

Repeat speakers

Our corpus contains numerous people who had more than one conversation: Of our 1456 unique participants, more than half had multiple conversations and approximately a third had three or more. This opens up many interesting questions, such as the variability in people's conversational behavior over time, the way conversation partners adapt to one another, the stability of the impressions that people make, and so forth. We regard this as a particularly unique and exciting aspect of the corpus.

Survey limitations

In quantifying auditory and visual conversational behavior, we sought to capture all of the information that current technology permits. In the future, better methods may become available to process and analyze audiovisual recordings. In contrast, the post-conversation survey necessarily covered a fixed (although broad) set of questions.

Intraconversational forecasting

We focused on three types of analysis across our findings: (i) patterns of low-level conversational features, such as turn-taking dynamics; (ii) high-level subjective outcomes, such as the link between conversation and well-being; and (iii) relations across the levels of the hierarchy, such as good conversationalists' use of vocal dynamism. One untapped area of inquiry is what we refer to as intraconversational forecasting, i.e., the way an intraconversational feature such as a speaker's tone of voice influences a listener's own reactions during that same turn or carries over into subsequent turns (consider the way laughter can be infectious). The premise of intraconversational forecasting, or the "flow of conversation" (*99*), is particularly intriguing as a subject for multimodal analysis; for example, a sudden shift in a listener's facial expression (e.g., a frown or a fading smile) may prompt the speaker to change the topic or soften their tone. The considerable range of potential questions, together with the complexity involved in parameterizing these kinds of temporal state inference models, makes intraconversational forecasting a particularly rich subject for future research.

Additional labeling

The more high-quality annotations (labels) that are added to the CANDOR corpus, the more useful its data will become. In its current state, the corpus already allows sophisticated machine learning models to be developed and trained. With more comprehensive labels, including better tracking of facial movements, together with enhanced emotional and semantic inferences, the potential applications become endless: Consider, for example, the way labeled datasets have been used recently in the computational study of linguistics to advance our understanding of decades-old concepts such as politeness (*100*) or the way multimodal deep learning architectures to detect emotion are increasingly used in conversational artificial intelligence. We are excited to see what other characteristics may be quantifiable, and we encourage scholars to make their annotations and analyses available to the broader community.

Additional detectors

While some software libraries today allow certain aspects of conversation, such as politeness (*101*), to be analyzed, we need many more (e.g., models that detect self-disclosure or high-accuracy laughter detectors). By releasing the entirety of the raw and processed corpus recordings, we anticipate that the corpus will grow together with advancing technologies.

Diversity of pairings

One strength of the corpus that is worth reiterating is its diversity of pairings, such as conversations between old and young and conversations across gender, race, and political orientation (see the "A qualitative glance at the corpus—Topical, relational, and demographic diversity" section). These pairings represent a major opportunity for research on intergroup contact.

By releasing the corpus to the public, it is our hope that other teams of researchers will push it to new heights: reprocessing, labeling, and extracting more features; analyzing people's stability and variability across repeat conversations; examining the way video-mediated conversation differs from face-to-face conversation; and pursuing answers to the large questions that remain outstanding. Perhaps other researchers will release their own complementary datasets of groups, friends, or work colleagues; of people with social anxiety; or of those talking across group divides. Ultimately, the expansion of this corpus, together with the release of more corpora, will allow accelerated progress toward a science of conversation.

Over the course of 2020, nearly 1500 people ranging in age from 19 to 66 were paired and engaged in recorded online video conversations. These recordings, which contain more than 7 million words across 850+ hours, together make up the CANDOR corpus, which we have introduced here. The wealth of linguistic, acoustic, visual, behavioral, and textual data that comprise this corpus allows researchers in a number of scientific disciplines to open fresh lines of inquiry into the most fundamental of human social activities: the spoken conversation.

We strongly encourage people to take the time to actually watch these recordings. We imagine that you will find, as we did, not only ideas for future research but also conversation's notable power to connect people. Despite the awkward small talk, the differing politics, and the understandable reticence—at least initially—on the part of strangers meeting for the first time, people nonetheless managed to come together, often with great kindness, grace, and understanding.

This orientation toward social connection is not only among the distinguishing features of our species but is also fundamental to the act of conversation itself. After all, conversation requires of its participants a remarkable degree of cognitive and social interdependence, from the joint construction of dialog to the inexorable search for a common ground and to the nuanced coordination of one's informational and relational goals. This was a joy to watch, although it was clear to us that the speakers themselves experienced the real joy, engaged as they were in the magic of building a shared experience through the spoken word. We should be grateful as scholars of human behavior that so much of this ancient ritual remains open to investigation.

Supplementary Materials

This PDF file includes:

Tables S1 to S6
Figs. S1 to S14
References

REFERENCES AND NOTES

1. H. H. Clark, *Arenas of Language Use* (University of Chicago Press, 1992).
2. N. J. Enfield, *How We Talk: The Inner Workings of Conversation* (Basic Books, 2017).
3. M. J. Pickering, S. Garrod, *Understanding Dialogue: Language Use and Social Interaction* (Cambridge Univ. Press, 2021).
4. H. Sacks, E. A. Schegloff, G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, in *Studies in the Organization of Conversational Interaction*, J. Schenkein, Ed. (Academic Press, 1978), pp. 7–55.
5. M. Tomasello, *Constructing a Language: A Usage-based Theory of Language Acquisition* (Harvard Univ. Press, 2003).
6. M. C. Bateson, Mother-infant exchanges: The epigenesis of conversational interaction. *Ann. N. Y. Acad. Sci.* **263**, 101–113 (1975).
7. C. Trevarthen, K. J. Aitken, Infant intersubjectivity: Research, theory, and clinical applications. *J. Child Psychol. Psychiatry* **42**, 3–48 (2001).
8. S. C. Levinson, J. Holler, The origin of human multi-modal communication. *Philos. Trans. R Soc. B Biol. Sci.* **369**, 20130302 (2014).
9. S. Pika, R. Wilkinson, K. H. Kendrick, S. C. Vernes, Taking turns: Bridging the gap between human and animal communication. *Proc. R. Soc. B* **285**, 20180598 (2018).
10. J. Henrich, *The Secret of Our Success* (Princeton Univ. Press, 2015).
11. E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, M. Tomasello, Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science* **317**, 1360–1366 (2007).
12. R. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1998).
13. R. I. Dunbar, Gossip in evolutionary perspective. *Rev. Gen. Psychol.* **8**, 100–110 (2004).
14. J. Holt-Lunstad, T. F. Robles, D. A. Sbarra, Advancing social connection as a public health priority in the United States. *Am. Psychol.* **72**, 517–530 (2017).
15. A. Milek, E. A. Butler, A. M. Tackman, D. M. Kaplan, C. L. Raison, D. A. Sbarra, S. Vazire, M. R. Mehl, "Eavesdropping on happiness" revisited: A pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality. *Psychol. Sci.* **29**, 1451–1462 (2018).
16. E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, A. Rudnicky, J. Williams, J. Pineau, M. Burtsev, J. Weston, The second conversational intelligence challenge (convai2), in *The NeurIPS'18 Competition* (Springer Cham, 2020), pp. 187–208.
17. A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, A. Pettigrue, Conversational AI: The science behind the alexa prize. arXiv:1801.03604 [cs.AI] (11 January 2018).
18. S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y. Boureau, J. Weston, Recipes for building an open-domain chatbot. arXiv:2004.13637 [cs.CL] (28 April 2020).
19. A. Anderson, M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. Mcallister, J. Miller, C. Sotillo, H. S. Thompson, R. Weinert, The HCRC map task corpus. *Lang. Speech* **34**, 351–366 (1991).
20. J. J. Godfrey, E. C. Holliman, J. McDaniel, SWITCHBOARD: Telephone speech corpus for research and development, in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (IEEE, 1992), pp. 517–520.
21. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv:1810.02508 [cs.CL] (5 October 2018).
22. S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku, Emotionlines: An emotion corpus of multi-party conversations. arXiv:1802.08379 [cs.CL] (23 February 2018).
23. N. Barros, E. Churamani, H. Lakomkin, H. Sequeira, A. Sutherland, S. Wermter, The OMG-emotion behavior dataset, in *Proceedings of the International Joint Conference on Neural Networks* (IEEE, 2018), pp. 1408–1414.
24. I. V. Serban, R. Lowe, P. Henderson, L. Charlin, J. Pineau, A survey of available corpora for building data-driven dialogue systems. arXiv:1512.05742 [cs.CL] (17 December 2015).
25. S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference. arXiv:1508.05326 [cs.CL] (21 August 2015).
26. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL] (11 October 2018).
27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
28. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL] (16 January 2013).
29. R. Munro, S. Bethard, V. Kuperman, V. Tzuyin Lai, R. Melnick, C. Potts, T. Schnoebelen, H. Tily, Crowdsourcing and language studies: The new generation of linguistic data, in *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk* (Association for Computational Linguistics, 2010), pp. 122–130.
30. The ManyBabies Consortium, Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* **3**, 24–52 (2020).

31. Many Primates, D. M. Altschul, M. J. Beran, M. Bohn, J. Call, S. DeTroy, S. J. Duguid, C. L. Egelkamp, C. Fichtel, J. Fischer, M. Flessert, D. Hanus, D. B. M. Haun, L. M. Haux, R. A. Hernandez-Aguilar, E. Herrmann, L. M. Hopper, M. Joly, F. Kano, S. Keupp, A. P. Melis, A. M. Rodrigo, S. R. Ross, A. Sánchez-Amaro, Y. Sato, V. Schmitt, M. K. Schweinfurth, A. M. Seed, D. Taylor, C. J. Völter, E. Warren, J. Watzek, Establishing an infrastructure for collaboration in primate cognition research. *PLOS ONE* **14**, e0223675 (2019).
32. N. A. Coles, J. K. Hamlin, L. L. Sullivan, T. H. Parker, D. Altschul, Build up big-team science. *Nature* **601**, 505–507 (2022).
33. T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, S. C. Levinson, Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10587–10592 (2009).
34. S. C. Levinson, Turn-taking in human communication—Origins and implications for language processing. *Trends Cogn. Sci.* **20**, 6–14 (2016).
35. M. Heldner, J. Edlund, Pauses, gaps and overlaps in conversations. *J. Phon.* **38**, 555–568 (2010).
36. J. P. de Ruiter, H. Mitterer, N. J. Enfield, Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* **82**, 515–535 (2006).
37. C. Riest, A. B. Jorschick, J. P. de Ruiter, Anticipation in turn-taking: Mechanisms and information sources. *Front. Psychol.* **6**, 89 (2015).
38. S. Bögels, F. Torreira, Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phon.* **52**, 46–57 (2015).
39. L. Magyari, Predictions in conversation, in *A Life in Cognition*: *Studies in Cognitive Science in Honor of Csaba Pléh*, J. Gervain, G. Csibra, K. Kovács, Eds. (Springer Cham, 2022), pp. 59–75.
40. S. C. Levinson, F. Torreira, Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* **6**, 731 (2015).
41. L. Ten Bosch, N. Oostdijk, L. Boves, On temporal aspects of turn taking in conversational dialogues. *Speech Commun.* **47**, 80–86 (2005).
42. R. E. Corps, B. Knudsen, A. S. Meyer, Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition* **223**, 105037 (2022).
43. N. G. MacLaren, F. J. Yammarino, S. D. Dionne, H. Sayama, M. D. Mumford, S. Connelly, R. W. Martin, T. J. Mulhearnc, E. M. Todd, A. Kulkarnid, Y. Caoa, G. A. Ruark, Testing the babble hypothesis: Speaking time predicts leader emergence in small groups. *Leadersh. Q.* **31**, 101409 (2020).
44. M. S. Mast, Dominance as expressed and inferred through speaking time: A meta-analysis. *Hum. Commun. Res.* **28**, 420–450 (2002).
45. A. Hepburn, G. B. Bolden, *Transcribing for Social Research* (Sage, 2017).
46. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
47. J. B. Bavelas, J. Gerwing, The listener as addressee in face-to-face dialogue. *Int. J. List.* **25**, 178–198 (2011).
48. R. Gardner, *When Listeners Talk* (John Benjamins Publishing Company, 2001).
49. J. B. Bavelas, L. Coates, T. Johnson, Listeners as co-narrators. *J. Pers. Soc. Psychol.* **79**, 941–952 (2000).
50. J. Tolins, J. E. F. Tree, Addressee backchannels steer narrative development. *J. Pragmat.* **70**, 152–164 (2014).
51. G. Jefferson, Caveat speaker: Preliminary notes on recipient topic-shift implicature. *Res. Lang. Soc. Interact.* **26**, 1–30 (1993).
52. G. Brown, G. Yule, *Discourse Analysis* (Cambridge Univ. Press, 1983).
53. H. P. Grice, Logic and conversation, in *Speech Acts* (Brill, 1975), pp. 41–58.
54. P. Brown, S. C. Levinson, *Politeness: Some Universals in Language Usage* (Cambridge Univ. Press, 1987), vol. 4.
55. S. Albert, J. P. de Ruiter, Repair: The interface between interaction and cognition. *Top. Cogn. Sci.* **10**, 279–313 (2018).
56. N. D. Duran, A. Paxton, R. Fusaroli, ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychol. Methods* **24**, 419–438 (2019).
57. C. Dideriksen, M. H. Christiansen, K. Tylén, M. Dingemanse, R. Fusaroli, Quantifying the interplay of conversational devices in building mutual understanding. PsyArXiv (12 October 2020).https://psyarxiv.com/a5r74/.
58. D. S. Berry, J. S. Hansen, Positive affect, negative affect, and social interaction. *J. Pers. Soc. Psychol.* **71**, 796–809 (1996).
59. L. A. Clark, D. Watson, Mood and the mundane: Relations between daily life events and self-reported mood. *J. Pers. Soc. Psychol.* **54**, 296–308 (1988).
60. L. C. Hawkley, J. T. Cacioppo, Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Ann. Behav. Med.* **40**, 218–227 (2010).
61. N. Epley, M. Kardas, X. Zhao, S. Atir, J. Schroeder, Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends Cogn. Sci.* **26**, 406–418 (2022).
62. N. Epley, J. Schroeder, Mistakenly seeking solitude. *J. Exp. Psychol. Gen.* **143**, 1980–1999 (2014).
63. J. Schroeder, D. Lyons, N. Epley, Hello, stranger? Pleasant conversations are preceded by concerns about starting one. *J. Exp. Psychol. Gen.* **151**, 1141–1153 (2021).
64. E. J. Boothby, G. Cooney, G. M. Sandstrom, M. S. Clark, The liking gap in conversations: Do people like us more than we think? *Psychol. Sci.* **29**, 1742–1756 (2018).
65. G. Cooney, E. J. Boothby, M. Lee, The thought gap after conversation: Underestimating the frequency of others' thoughts about us. *J. Exp. Psychol. Gen.* **151**, 1069–1088 (2022).
66. A. Mastroianni, G. Cooney, E. J. Boothby, A. G. Reece, The liking gap in groups and teams. *Organ. Behav. Hum. Decis. Process.* **162**, 109–122 (2021).
67. E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. C. LeBeaumont, T. Wheatley, Fast response times signal social connection in conversation. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2116915119 (2022).
68. B. L. Smith, B. L. Brown, W. J. Strong, A. C. Rencher, Effects of speech rate on personality perception. *Lang. Speech* **18**, 145–152 (1975).
69. N. Miller, G. Maruyama, R. J. Beaber, K. Valone, Speed of speech and persuasion. *J. Pers. Soc. Psychol.* **34**, 615–624 (1976).
70. N. A. Murphy, J. A. Hall, C. R. Colvin, Accurate intelligence assessments in social interactions: Mediators and gender effects. *J. Pers.* **71**, 465–493 (2003).
71. M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in *International Conference on Machine Learning* (PMLR, 2015), pp. 957–966.
72. K. Song, X. Tan, T. Qin, J. Lu, T. Y. Liu, MPNet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.* **33**, 16857–16867 (2020).
73. N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv:1908.10084 [cs.CL] (27 August 2019).
74. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 [cs.CL] (2019).
75. M. Florentine, Loudness, in *Loudness*, M. Florentine, A. N. Popper, R. R. Fay, Eds. (Springer, 2011), pp. 1–15.
76. L. E. Marks, M. Florentine, Measurement of loudness, part I: Methods, problems, and pitfalls, in *Loudness*, M. Florentine, A. N. Popper, R. R. Fay, Eds. (Springer, 2011), pp. 17–56.
77. R. A. Page, J. L. Balloun, The effect of voice volume on the perception of personality. *J. Soc. Psychol.* **105**, 65–72 (1978).
78. A. C. Weidman, J. L. Tracy, Picking up good vibrations: Uncovering the content of distinct positive emotion subjective experience. *Emotion* **20**, 1311–1331 (2020).
79. E. Diener, R. J. Larsen, S. Levine, R. A. Emmons, Intensity and frequency: Dimensions underlying positive and negative affect. *J. Pers. Soc. Psychol.* **48**, 1253–1265 (1985).
80. R. Reisenzein, Pleasure-arousal theory and the intensity of emotions. *J. Pers. Soc. Psychol.* **67**, 525–539 (1994).
81. J. Sonnemans, N. H. Frijda, The structure of subjective emotional intensity. *Cogn. Emot.* **8**, 329–350 (1994).
82. S. R. Livingstone, F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* **13**, e0196391 (2018).
83. D. E. King, Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009).
84. A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**, 18–31 (2017).
85. L. F. Barrett, Solving the emotion paradox: Categorization and the experience of emotion. *Pers. Soc. Psychol. Rev.* **10**, 20–46 (2006).
86. U. Hess, C. Blaison, K. Kafetsios, Judging facial emotion expressions in context: The influence of culture and self-construal orientation. *J. Nonverbal Behav.* **40**, 55–64 (2016).
87. T. Stivers, J. Sidnell, *The Handbook of Conversation Analysis* (John Wiley & Sons, 2012).
88. E. Stokoe, *Talk: The Science of Conversation* (Hachette UK, 2018).
89. D. Jurafsky, D. Tolinsky, R. Ranganath, D. McFarland, Extracting social meaning: Identifying interactional style in spoken conversation, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2009), pp. 638–664
90. J. F. Dovidio, K. Kawakami, S. L. Gaertner, Implicit and explicit prejudice and interracial interaction. *J. Pers. Soc. Psychol.* **82**, 62–68 (2002).
91. M. Sen, O. Wasow, Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annu. Rev. Polit. Sci.* **19**, 499–522 (2016).
92. J. Henrich, S. J. Heine, A. Norenzayan, Most people are not WEIRD. *Nature* **466**, 29–29 (2010).
93. G. Cooney, A. Mastroianni, N. Abi-Esber, A. W. Brooks, The many minds problem: Disclosure in dyadic versus group conversation. *Curr. Opin. Psychol.* **31**, 22–27 (2020).
94. R. L. Moreland, Are dyads really groups? *Small Group Res.* **41**, 251–267 (2010).

95. T. Stivers, Is conversation built for two? The partitioning of social interaction. *Res. Lang. Soc. Interact.* **54**, 1–19 (2021).
96. E. Jolly, L. J. Chang, Gossip drives vicarious learning and facilitates social connection. *Curr. Biol.* **31**, 2539–2549.e6 (2021).
97. K. Greene, V. J. Derlega, A. Mathews, Self-disclosure in personal relations, in *The Cambridge Handbook of Personal Relations*, A. L. Vangelisti, D. Perlman, Eds. (Cambridge Univ. Press, 2006), pp. 409–427.
98. D. Wilkes-Gibbs, H. H. Clark, Coordinating beliefs in conversation. *J. Mem. Lang.* **31**, 183–194 (1992).
99. D. Knox, C. Lucas, A dynamic model of speech for the social sciences. *Am. Polit. Sci. Rev.* **115**, 649–666 (2021).
100. C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, A computational approach to politeness with application to social factors. arXiv:1306.6078 [cs.CL] (25 June 2013).
101. M. Yeomans, A. Kantor, D. Tingley, The politeness package: Detecting politeness in natural language. *R Journal* **10**, 489–502 (2018).
102. Y. Jadoul, B. Thompson, B. De Boer, Introducing parselmouth: A python interface to praat. *J. Phon.* **71**, 1–15 (2018).
103. B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in *Proceedings of the 14th Python in Science Conference* (2015), vol. 8, pp. 18–25.
104. P. Mermelstein, Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* **116**, 374–388 (1976).
105. M. Farrús, J. Hernando, P Ejarque, Jitter and shimmer measurements for speaker recognition, in *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (International Speech Communication Association, 2007), pp. 778–781
106. P. Boersma, D. Weenink, Praat: Doing phonetics by computer [Computer program], version 6.2.08 (2022);www.praat.org/.
107. SPTK Working Group, Speech Signal Processing Toolkit (SPTK) (2017);http://sp-tk.sourceforge.net.
108. N. Dehak, P. Dumouchel, P. Kenny, Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **15**, 2095–2103 (2007).
109. J. Howard, R. Thomas, S. Gugger, fastai. GitHub (2018).
110. S. Dutta, S. Datta, A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics* **72**, 432–440 (2016).
111. C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, A. Galbarczyk, G. Jasienska, C. T. Ross, M. Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, A. Schachner, T. A. Vardy, Q. D. Atkinson, A. Salenius, J. Andelin, J. Antfolk, P. Madhivanan, A. Siddaiah, C. D. Placek, G. Deniz Salali, S. Keestra, M. Singh, S. A. Collins, J. Q. Patton, C. Scaff, J. Stieglitz, S. Ccari Cutipa, C. Moya, R. R. Sagar, M. Anyawire, A. Mabulla, B. M. Wood, M. M. Krasnow, S. A. Mehr, Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **6**, 1545–1556 (2022).
112. C. Yale, A. B. Forsythe, Winsorized regression. *Technometrics* **18**, 291–300 (1976).
113. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).

Acknowledgments: We extend particular thanks to all of the participants in the corpus who agreed to make their conversations public. We also wish to thank BetterUp Inc., particularly A. Robichaux and G. Kellerman, for funding this research and for willingness to share the data collected for research among the wider scientific community. We also thank the entire team at DrivenData for considerable expertise in data science and machine intelligence, which made this undertaking possible. **Funding:** D.K. acknowledges financial support through the National Science Foundation (award #2120087 through the Methodology, Measurement, and Statistics Program). This research was made possible in part by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors. G.C. wishes to thank the Wharton Behavioral Lab for assistance running the pilot study and Wharton's OID Department for financial support. **Author contributions:** Apart from A.R. and G.C., the remaining authors are listed alphabetically. A.R., P.B., C.F., C.C., and T.G. built the pipeline to recruit participants and process conversations. P.B., C.F., C.C., and T.G. developed turn-taking algorithms and feature extraction methods. A.R. and P.B. designed and oversaw the final public release version of the dataset. G.C. and A.R. designed the survey and provided the overall direction for the analysis. A.L. performed the analyses for the sections related to turn-taking, with support from G.C. and A.R. G.C. and A.L. performed the analyses related to well-being. D.K. and G.C. designed the section "Characteristics of good and bad conversationalists," with support with A.R; D.K. also performed the analysesand wrote the Supplementary Materials, with support from G.C., A.R., A.L., and P.B. D.K., G.C., and A.R. are responsible for the section "A qualitative glance at the corpus—Topical, relational, and demographic diversity." A.L. and D.K. created the pipeline to process the raw data into analysis-ready data, with support from A.R. S.M. and G.C. led the design and creation of the final figures with support from A.L. and A.R. B.D. performed the qualitative review. G.C. drafted the initial manuscript. A.R. and D.K. provided edits to the initial manuscript. G.C. produced the final revised manuscript. G.C. handled the peer review process, with support from D.K. and A.R. **Competing interests:** At the time this project was conducted, authors with BetterUp affiliations were paid employees of BetterUp Inc.; DrivenData Inc., authors with a DrivenData affiliation, and G. Cooney were paid consultants at BetterUp Inc. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. We are also publicly releasing the full CANDOR corpus for the broader scientific community's use. Data are available for access via registration here: https://betterup-data-requests.herokuapp.com/. Researchers can choose to download "raw media" (i.e., raw video/audio files for each speaker), "processed media" (i.e., processed, synced, and aligned video/audio files for each conversation), and/or "no media" (i.e., just the survey data). Materials, code, and important links are available here: https://osf.io/fbsgh/. This repository contains preparation and processing scripts (i.e., Python and R scripts that take the downloaded data and output the final analysis-ready files); analysis scripts (i.e., R code to reproduce all of the analyses that appear in the manuscript); and important supplementary files, such as the Data Dictionary (i.e., a spreadsheet that describes all of the survey variables and other extracted features) and the Qualitative Review (i.e., a spreadsheet that contains the results of our qualitative review of all conversations).

Submitted 13 October 2022
Accepted 2 March 2023
Published 31 March 2023
10.1126/sciadv.adf3197