



Suspicious Minds: the Problem of Trust and Conversational Agents

Jonas Ivarsson*¹  & Oskar Lindwall¹

*¹Department of Applied IT, University of Gothenburg, Gothenburg, Sweden
(E-mail: jonas.ivarsson@gu.se; E-mail: oskar.lindwall@gu.se)

Accepted: 27 February 2023

Abstract. In recent years, the field of natural language processing has seen substantial developments, resulting in powerful voice-based interactive services. The quality of the voice and interactivity are sometimes so good that the artificial can no longer be differentiated from real persons. Thus, discerning whether an interactional partner is a human or an artificial agent is no longer merely a theoretical question but a practical problem society faces. Consequently, the ‘Turing test’ has moved from the laboratory into the wild. The passage from the theoretical to the practical domain also accentuates understanding as a topic of continued inquiry. When interactions are successful but the artificial agent has not been identified as such, can it also be said that the interlocutors have understood each other? In what ways does understanding figure in real-world human–computer interactions? Based on empirical observations, this study shows how we need two parallel conceptions of understanding to address these questions. By departing from ethnomethodology and conversation analysis, we illustrate how parties in a conversation regularly deploy two forms of analysis (categorical and sequential) to understand their interactional partners. The interplay between these forms of analysis shapes the developing sense of interactional exchanges and is crucial for established relations. Furthermore, outside of experimental settings, any problems in identifying and categorizing an interactional partner raise concerns regarding trust and suspicion. When suspicion is roused, shared understanding is disrupted. Therefore, this study concludes that the proliferation of conversational systems, fueled by artificial intelligence, may have unintended consequences, including impacts on human–human interactions.

Keywords: Conversation, Human–computer interaction, Natural language processing, Trust, Understanding

1 Introduction

*We can't go on together
With suspicious minds
And we can't build our dreams
On suspicious minds*

Mark James wrote the song, which depicts a couple caught in a mistrusting and dysfunctional relationship, from which we borrowed our title. As with any good pop song, much meaning is contained in a few select words. An examination of the first line of the chorus allows us to elaborate on the aims of this essay. We will expand the scope of the implied relationship from lovers to persons interacting with digital agents or chatbots. By shifting the focus to these parahuman relations, the paper aims to highlight the profound philosophical question of what acting in concert with others means—to ‘go on together’, as the song describes it. The keystone for the lyrics then picks up on suspicion as the very obstacle to a shared project.

The ability to ‘go on’ has been connected to understanding, most notably by Wittgenstein (1953). Understanding one another is perhaps not a prerequisite for cooperation, but it is deeply intertwined with the effectiveness of concerted action and the entailed intersubjectivity structures of our sense of reality (Schutz, 1976). Stahl (2016) argued that the question of how people can understand each other is a foundational issue for the social sciences with particular relevance to CSCW, as studies in the field continuously contribute analyses of how intersubjectivity is established in specific work practices.

However, any detailed examination of understanding or intersubjectivity is bound to confront a wealth of thorny theoretical and methodological issues. In pursuing our goal, we draw on the works of Wittgenstein, Garfinkel, and Sacks to align ourselves with an empirical and conceptual approach to CSCW that Randall et al. (2021) have characterized by its ‘careful examination of the nature of concept use and some scepticism about representational views of the mind’ (p. 191). As we aim to demonstrate further, this carefulness becomes especially important when bringing agentic technologies within the analytic scope.

Our purpose in what follows is to address the question of what it means to act together, to cooperate, in a landscape potentially permeated by fake or artificial actors—that is, in what sense is interaction with the artificial itself contrived as a mere simulacrum? As articulated by McDermott, ‘When we can talk to machines, will we understand each other?’ (McDermott, 2007, p. 1183). More specifically, we aim to expand prior work in CSCW by examining circumstances in which suspicion, deception, and disbelief are occasioned as natural phenomena. We will also inspect practices deployed to maintain the distinction between human and artificial interlocutors and the conditions under which this separation becomes strained. Central to our argument is the separation between two notions of understanding: interactional and dispositional. While the production of a relevant next turn is a constitutive feature of understanding-in-interaction, the performance of such a turn is not sufficient for ascribing understanding to conversational agents as a dispositional property. After having described three separate phone calls—involving a bot

Suspicious Minds: the Problem of Trust and Conversational Agents

calling a human, a human calling a human, and a human calling a bot—we will conclude by discussing certain implications of these observations for CSCW research on conversational agents.

2 Understanding Human–AI Interaction

The literature addressing human–AI interaction, including conversational systems, is increasing rapidly, with special issues devoted to topics such as conversational user interfaces and interactions (Landay et al., 2019), transparent human–agent communications (Chen, 2022), and the question of how to design and manage human–AI interactions (Abedin et al., 2022). The different ways that “understanding” or “competence” play into the exchange are a recurring theme, and principles have been proposed to support the design of effective voice-based human–machine interaction. For instance, users are understood to fail to engage with speech-enabled devices when they use humanlike voices that misrepresent their true capabilities (Moore, 2017a). Others have addressed the skills on the opposite side of the human–machine relation. Da Silva et al. (2022) investigated the use of Google Assistant by comparing people with varying literacy levels. Their experiment showed that Google Assistant had difficulties ‘understanding the intention’ (p. 12) of illiterate users as the technology struggled to parse the grammatical structure of sentences produced in an ad hoc manner.

In addition to experimental designs, the proliferation of conversational technologies in peoples’ phones, homes, and cars enables studies of naturally occurring talk. Porcheron et al. (2018) studied voice user interfaces (Amazon Echo) embedded in participants’ homes and examined how the devices were made a part of everyday conversations. The authors take a strong stance against anthropomorphic characterizations and reject the notion that such devices and interfaces are conversational in nature. The claim is that “‘conversational interaction’ is a misnomer for this kind of human-computer interaction, and confuses interaction with a device *within* conversation with an *actual* conversation.’ (Porcheron et al., 2018, p. 640, emphasis in original). This critical appraisal of the positioning of agentic technologies in HCI research, particularly regarding the notion of ‘interaction’ (Reeves and Beck, 2019), has been carried further by demonstrations of the organized ways conversational AI systems are embedded into everyday action. Reeves and Porcheron (2022) argue for reframing people’s use of interactive AI technologies, not as interactions, but as technology *regulated* within a social organization.

Common to much of this literature is the ‘knowing’ character of the studied episodes of use. That is, users are regularly aware of the fact that assistive technologies are synthetic and machine-based. Our aim here is to consider situations in which that sense of knowing or awareness is under duress. In this way,

the paper can expand the discussion in CSCW and related fields to appraise the accomplishment of suspicion and deception, as well as ‘playing along’ or suspension of disbelief when doing interactions with conversational AIs. The scenario we are targeting is not new—it was formulated in the earliest days of computer science.

3 What Does it Mean ‘to Understand’? Two Uses of the Term

In 1950, Alan Turing formulated his ‘imitation game’, which later became known as the ‘Turing Test’. This thought experiment was modeled on a parlor game and was cast like a game of deception with three parties: a machine (A), a human (B), and an interrogator (C). Armed with suspicion—the knowledge that one of the conversational partners is a machine—the interrogator’s task was to reveal A and B’s true identities through a series of questions and answers mediated via a teleprinter communication arrangement. Turing proposed replacing the question ‘Can machines think?’ with this test. If the interrogator regularly fails to differentiate the machine from the human through questions and answers, we should ascribe a measure of intelligence to the machine.

A critique of the Turing Test, formulated by Graham Button et al. (1995), begins with Wittgenstein’s observation that the proposition ‘But a machine surely cannot think’ (Wittgenstein, 1953, paragraph 360) is not an empirical statement but a conceptual one. For the Turing test to work, we must accept that thinking is demonstrated if the computer can do the same things as intelligent people. However, a machine might pass the test in the sense of replicating what intelligent humans would do in a comparable situation, without this suggesting that the machine thinks or that intelligence should be ascribed to its performance. As Button et al. (1995) have noted, we might very well be unable to tell the difference between a live and recorded musical performance when played in the next room. If we knew the source of the music, however, we would never dream of talking about the musical skills of the recording device as we would of the musician. Our inability to separate the two simply means that we can be deceived to hear one thing as another. Through its design, the Turing test hides the apparent difference between humans and machines, but in doing so, it also erodes the grounds on which concepts such as thinking or intelligence are formulated. The design of the test turns the question ‘Can machines think?’ into a proposition to be empirically settled; at the same time, it ‘presupposes the very things which the test is meant to demonstrate’ (Button et al., 1995, p. 142). Therefore, Button et al. contended that ‘[t]he logic of the Turing test is to ask this question: overlooking the fact that this is a machine, can a machine think?’ (ibid.)

Wittgenstein (1953) argued that the grammar of understanding is linked to action and performance and belongs to a different realm than mental states,

Suspicious Minds: the Problem of Trust and Conversational Agents

experiences, or processes. Given that the Turing Test exclusively focuses on performance and outcomes, not what goes on inside humans or machines, it could be argued that the test expresses a conception of understanding similar to Wittgenstein's position. However, according to Button et al., this interpretation would be a mistake, resulting from a 'superficial reading of Wittgenstein' (p. 145). It is a reading that results in an 'inappropriately narrow focus upon some features of the performance itself, disassociating it from the background circumstances against which the question of whether one performance is comparable with, and shows the same as, another, must be posed' (p. 146). Two actions might be impossible to differentiate solely based on their physical characteristics, but this does not mean they are equal. As the authors pointed out, a forged banknote is different from an official and legitimate currency, regardless of the quality of the forgery. The difference between the two cannot be reduced to their physical characteristics. As Button et al. further argued, someone might display an understanding of arithmetic by using pen and paper, an abacus, or by performing calculations in their head. However, if they provided correct answers by repeating what someone else told them or by reading from a note, we would not count that as displaying an understanding of arithmetic. That we ignore what is going on in the head 'does not equate to a conviction that the manner in which a performance is produced is itself an irrelevance to the judgment as to whether it displays understanding' (ibid.).

Although we agree with this critique of the Turing test, there are reasons for identifying other distinguishing features in the grammar of understanding. Turing's central thesis was to put aside our normal use of the words 'machines' and 'thinking' and propose an alternate formulation of the problem. Through this operationalization, he also placed blinders on the variety of usages that we can observe in common parlance. As outlined in the critique by Button et al. (1995), understanding appears *ascribable* to certain activities but not to others—on conceptual grounds and typically with reference to the agents involved. When understanding is discussed in this sense, it qualifies as a dispositional statement. Ryle (1949) described the logic of dispositional concepts as follows:

When we describe glass as brittle, or sugar as soluble, we are using dispositional concepts, the logical force of which is this. The brittleness of glass does not consist in the fact that it is at a given moment actually being shattered. It may be brittle without ever being shattered. To say that it is brittle is to say that if it ever is, or ever had been, struck or strained, it would fly, or have flown, into fragments. (Ryle, 1949, p. 43)

According to Ryle (1949), when someone is said to understand Swedish, this should not be seen as a statement describing an act or event. We can observe two

parties in a conversation spoken entirely in Swedish and conclude that they both understand the language. While we are not in error in making this inference, the disposition or complex of dispositions cannot be observed or recorded, since it is a factor of the wrong logical type.

In contrast to the dispositional use of understanding, there is a very different usage of the term in which understanding is found entirely in action. This secondary usage turns understanding into a *constitutive feature* of sequences of action, where it becomes observable at every point.

Moerman and Sacks (1988) discussed how the turn-taking system used in a conversation requires that members constantly monitor the utterances of others to decide whether and when to speak next. As they pointed out, ‘any intended next speaker must work on understanding the current utterance to know what it will take for that utterance to be completed’ (p. 183). Consequently, if a member who has been selected to talk does not do so, this is taken ‘as evidence of failing to understand what has been said’ (ibid.). To be an active party in a conversation, members must be involved in the local and public task of demonstrating understanding. As Moerman and Sacks put it,

[U]nderstanding matters as a natural phenomenon in that conversational sequencing is built in such a way as to require that participants must continually, there and then – without recourse to follow up tests, mutual examination of memoirs, surprise quizzes and other ways of checking on understanding – demonstrate to one another that they understood or failed to understand the talk they are party to (Moerman and Sacks, 1988, p. 185).

The argument is not that understanding is restricted to conversational sequencing, but rather that conversational understanding, or meaning, cannot be divorced from the local production of talk-in-interaction. Although this resonates with a Wittgensteinian position on action, understanding, and meaning (Mair and Sharrock, 2021), it might be taken to indicate that voice assistants display an understanding by, for instance, answering a question. The conversation analytic notion of progressivity, to keep interactions moving forward, has also been highlighted as central for voice interface design (Fischer et al., 2019). Here, the distinction between constitutive and dispositional usages of understanding is relevant. According to Moerman and Sacks (1988), producing an appropriate utterance on time is the central task and display of understanding in conversation. Without this constitutive feature of talk-in-interaction, the conversation could not progress—at least not in a recognizable way as a mundane conversation between two competent adults. However, this does not suggest that we would necessarily ascribe understanding to conversational agents as dispositional properties, even in the face of them being competent interactional partners.

Suspicious Minds: the Problem of Trust and Conversational Agents

In summary, this section outlines two senses of what it means to understand: understanding characterized as dispositional statements and the alternative treatment of understanding as a constitutive feature of action. Nevertheless, we do not wish to settle the matter as a conceptual argument. Instead, we aim to further examine the extent to which both aspects are at play in real-world interactions. By observing lay society members' analyses and categorizations, we can study the ongoing achievements of trust and suspicion as issues connected to joint understanding.

4 Sequential and Categorical Analysis

In the subsequent sections, as we turn to the ways in which trust and suspicion play out in actual situations of human (and non-human) conduct, we draw on the findings and analytic mentality of conversation analysis (CA). One premise of CA is that the work of analysis is not restricted to that of professionals. As noted by Watson (1994), the name conversation analysis was, in the first instance, 'designed to designate a topic, namely interlocutors' own conjoint and culturally methodic analyses of their conversational actions, rather than designating an analytically privileged "special technique"' (p. 178). Moerman and Sacks' (1988) discussion of how an utterance "is *sequentially analyzable* for its possible completion" is not about how they, as sociologists, analyze the turn, but how parties to the conversation do that. As members of a conversation, we continually analyze the conversational turns of other members, thereby finding out when to talk and what an appropriate response might be. Through our responses, moreover, we show that we have analyzed the prior action in a certain way. If the utterance 'What are you doing tomorrow?' is met by, 'I'm sorry, but I am busy,' the second turn shows how the first is understood: the first utterance is not merely a question, but a 'pre-invitation'; this invitation has the second party as a relevant recipient. That the second utterance displays an understanding of the first also offers the possibility of repairing potential misunderstandings in the next turn: for instance, by saying 'No, I was just asking,' or 'Sorry, I wasn't talking to you.' The fact that members of a conversation in this way display how they come to understand each other's actions on a turn-by-turn basis then provides the professional 'conversational analyst' material for their second-order analysis.

As an academic field, CA is known for its detailed interest in the *sequential* organization of conversation, including the organization of turns, how successive turns are produced to be coherent with prior ones, and how troubles are monitored and handled (Schegloff, 2007). As the founder of CA, Sacks, (1992a, 1992b) was also interested in how members use language to *categorize* and classify the world. What distinguishes CA from other approaches to categorization in the social sciences is its interest in the local, reflexive, and occasioned character of category work. Categorizations are not only

investigated for how they distinguish people and events, but also for the ways in which they are constitutive of the local circumstances of production. In conversation analytic literature, sequential analysis and categorial analysis are sometimes kept distinct. As pointed out by Watson, however, members (and professional analysts) inevitably make sense of sequential aspects in relation to membership categorization procedures (Wovk and Carlin, 2004). In his own words, ‘The practices of “categorisation” and “sequencing” are not only inseparable but reciprocally – constituting and reciprocally – shaping. This reciprocity is, indeed, one among many of the “essential reflexivities” of conversational organisation’ (Watson, 2015, p. 35).

Regarding Sacks’ (1992b) work on telephone conversations, Watson (2015) noted that the categorizations ‘caller’ and ‘called’ shape conversations by organizing ‘the distribution of oriented-to sequential rights and obligations’ (p. 33). It is the caller and not the called party who has the right and obligation to formulate the reason for the call and initiate its closing. Such categorizations are further crucial in shaping the grasped meaning of actions that take place in interaction. If the identity of a conversational partner is tied in with a specific professional category, utterances can take on the character as evidence of professional conduct. However, if this categorization can no longer be maintained for some reason, the re-categorization can shift the meaning of an exchange retroactively: what was heard before is now understood in a new way.

When identifications are made by way of categories, they are not necessarily definite or settled. On occasion, the identity may be reconsidered or reevaluated based on additional information, leading to identity redocumentation. This, in turn, might lead to an attitude of suspicion and disbelief. With reference to Sacks’ (1972) work on police assessment of moral character, Watson noted ‘how trusted identities of persons – identities naïvely presented and received – come to be re-documented by the police “on the spot” as suspicious, possibly criminal’ (Watson, 2009, p. 494). These re-categorizations are good examples of the forms of laypersons’ sociological descriptions that could be examined for how the notions of trust and suspicion surface as topics; for example, ‘the re-categorization of a person from “businessman” to “drug dealer” is a prime instance of re-description – one that, plausibly, involves a withdrawal or reduction of trust’ (Watson, 2009, p. 495).

In bringing this perspective to bear on interactions with artificial conversational systems, we want to raise several questions: How are identifications made in situ, and in what ways will locally occasioned suspicion inform such identifications? Furthermore, how may interactional incongruities trigger identity redocumentation and, consequently, modify any sense of shared understanding?

Suspicious Minds: the Problem of Trust and Conversational Agents

What are the interactional consequences of operating with trust versus suspicion in dealing with other parties, regardless of their true identities?

5 The Artificial in Real Life

To pursue these questions, we will describe and analyze three separate phone calls. The conversations include a bot calling a human, a human calling another human, and finally, a human calling a bot. As classifications of the participating interactants, these categories are not necessarily given from the outset. On the contrary, they often stand as interactional accomplishments as the result of practical reasoning. The variations in the positions nevertheless allow for a series of observations.

The first situation will describe how deception can go unnoticed, and that unremarkable interaction with the artificial is now a technical possibility. The intent of this instance is for a bot to sound as human as possible for the purpose of completing a scheduling task.

The second call shares many similarities with the first. The setting is comparable, with the exception that the two parties in the conversation are both human. In this case, the human masquerades as a bot designed to sound as human as possible to complete a task. This additional layer of interpretation makes the call relevant in the context of our problem. That is, how an orderly conversation can be heard retroactively as contrived. We will discuss the evidence and reasons for this hearing-with-suspicion.

Finally, we will introduce a different example where the caller unknowingly gets forwarded to a limited interactivity bot designed with the intention of keeping the conversation going for as long as possible. We will show how this is made feasible through the positioning of the called as someone who can be excused for repeatedly displaying problems of hearing and understanding, and how the bot is specifically designed to ward off (or delay) suspicion despite its severely incoherent performance.

5.1 Silicon Valley Speak

Today, technology companies are working hard to develop advanced forms of natural language processing. Systems such as Alexa, Siri, and Google Assistant have become common in many households and mobile devices. In 2018, Google showcased a system named Duplex at its developer conference. This system was introduced as a service that could help users contact small businesses that would take bookings only by phone. The following conversation¹ was then played for

¹ <https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html>

the attending audience. In the setup, we are led to understand that the ‘caller’ is, in fact, Google’s artificial agent:

Extract 1. Google Duplex Haircut appointment²

01 CALLED: Hello how can I help you?
02 CALLER: Hi↑ I’m calling to book a woman’s haircut for a
03 client. uhm: I’m looking for something on may thi↑rd
04 (1.2)
05 CALLED: Sure, gimme o::↑ne second.
06 CALLER: mhm↑:
07 (2.7)
08 CALLED: sure what time are you looking for around?
09 (0.9)
10 CALLER: at twelve pm
11 CALLED: we do not have a twelve pm available↑ the closest
12 we have to that↑, is a one fifteen.
13 (0.8)
14 CALLER: do you have anything between ten am and e:h twelve pm
15 CALLED: depending on what service, she would like what service
16 is she looking for?
17 (0.8)
18 CALLER: just a woman’s haircut for now↑
19 CALLED: okay we have a ten o’clo:ck↑
20 (0.8)
21 CALLER: ten am is fine
22 CALLED: okay what’s her first name?
23 (1.0)
24 CALLER: the first name is Lisa↑
25 CALLED: okay perfect so I will see Lisa at ten o’clock
26 on may third.
27 (0.6)
28 CALLER: okay great thanks
29 CALLED: great have a great day bye

In this showcase of their new conversational technology, it is reasonable to suspect that Google selected a conversation that accentuates its capabilities in favor of

² We have tried to keep the transcriptions as accessible as possible. For analytic reasons, some details have been marked with the customary symbols used within conversation analysis as developed by Gail Jefferson (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation Analysis. Studies from the first generation* (pp. 13–31). John Benjamins. (0.0) Numbers in parentheses indicate elapsed time by tenths of seconds.:: Colons indicate prolongation of the immediately prior sound. The longer the colon row, the longer the prolongation. ↑ Upwards arrows indicate shifts into especially high pitch. ..? Punctuation markers are used to indicate ‘the usual’ intonation.

Suspicious Minds: the Problem of Trust and Conversational Agents

showing the world moments in which the technology may struggle. In this respect, the conversation exhibited is also unremarkable and ordinary. Beyond the setting of the call and the opening declaration expressing the intent of making an appointment on behalf of a client, there is no further discussion about the interacting parties' roles or motives. They transact their business in an orderly fashion, with booked appointment as a result. For the called, one could argue that the exchange is doing only that. Given the purpose of finding and deciding on a time for a haircut, the understanding produced in this interaction seems sufficient. As far as we can hear, there are no evident reactions from the hair salon side to suggest that this call is strange in any way.

As soon as the conversation was replayed before an audience of developers, some character changes were observed. Suddenly, the features of the talk produced by the caller became remarkable. The recurrent use of an upward pitch at the end of many sentences is difficult to grasp in the transcript. This variant of English is sometimes named 'high rising terminal' ('upspeak' or 'uptalk') and can be associated with the subcultural stereotype of 'valley girls'. In line 6, the response 'mhm' is fitted with this expressed upward inflection (marked by arrows in the transcript). When replayed at the Google conference, this moment is met with a burst of roaring laughter from the audience. To understand why this juncture was suddenly seen as a sensational event, we can zoom in on the exchange around this point.

In line 5, the taker of the call projects a momentary lapse in the conversation, thereby inviting the recipient to hold off talking where she might otherwise start (Schegloff, 1982). To the interactants, the response 'mhm' acknowledges this request. From a production perspective on understanding, this short reply provides the recipient with just enough evidence that she has been heard and understood. Talk will be withheld, and we see a longer (2.7 s) pause ensuing, which further corroborates this understanding as an interactional achievement.

For adult English speakers, events like these are so commonplace that they rarely become noticeable. For the audience at the conference, by contrast, seeing, perhaps for the first time in their lives, a computer managing such intricacies of conversation, the occasion becomes marked with celebratory amusement, hence the laughter.

Another aspect of the talk can be observed in the Valspeak accent. Ending declarative statements with a rising-pitch intonation poses a theoretical risk that they will be heard as questions. One could reason that this would be a problematic design for a conversational technology supposedly attempting to minimize problems of understanding. However, by making the digital assistant sound as natural as possible, people can form rapid judgments about their interactional partner's characteristics through their voices (Cambre and Kulkarni, 2019). In this way, the vocal dressing up of the conversational agent as a young Californian woman talking to another Californian woman could be seen as misdirection in the stage magic presented. In the cultural setting where it was deployed, the accent would help make the agent hearable (classifiable) as an ordinary speaker of English. It could be seen as one method

for not arousing any suspicion about the caller being a bot. Through this vocal dress, Google would take the imitation game (Turing, 1950) to the next level.

5.2 Hearing with Suspicion

When Google's new service became available, technology journalists and other interested citizens began to explore its significance. What ensued was something that we would like to describe as a *laic interest* in talk's production features. People who typically would not conduct formal investigations into the minutiae of spoken interaction—that is, amateurs, or enthusiasts—now started to produce and discuss such analyses to discover who was or was not an AI. With the general public now roused to suspicion on the grounds of possible deception, people began to try to identify artificial deceivers. Turing's imitation game escaped the laboratories and ventured out into the larger world. Some would stage their own dinner reservation recordings with the Duplex service and make them available for further analysis, such as the following from Venture Beat (Wiggers, 2018):

Extract 2. Café Prague Dinner Reservation

01 CALLED: Hello café Prague↑
02 (1.5)
03 CALLER: Hi: I'm calling for a client, I'm calling from Google
04 so the call may be recorded, (.) can I book a table
05 for tomorrow please↑
06 CALLED: Yes at what time would you like to reserve
07 (1.5)
08 CALLER: At seven pm↑
09 CALLED: O↑key perfect how many people↑
10 (1.7)
11 CALLER: It's for two people
12 CALLED: O↑key, two people tomorrow at seven pm. What's your
13 name↑
14 (1.6)
15 CALLER: It's- her first name is Anna↑
16 CALLED: Anna, okey thank you.
17 (1.9)
18 CALLER: Okey
19 CALLED: I see you tomorrow thank you
20 (1.2)
21 CALLER: Thank you and thank you very much, thank you bye bye
22 CALLED: Bye bye

Suspicious Minds: the Problem of Trust and Conversational Agents

In the actual video,³ the staged character of this call to the restaurant is evident. The call-taker is clearly aware of being recorded for the sake of documenting Google's new service. At this instance, however, we are less interested in her first-order analysis and focus instead on a battery of secondary analyses afforded by the recording as such. When the video clip was uploaded online, people began scrutinizing the minute details of how language was used there to find evidence that would give away the caller as a robot (*Google Duplex: AI assistant makes a restaurant reservation*, 2018). One commenter pointed to the 1.5-second pause at the beginning of the call. Another wrote that they could easily spot that it was a bot, given the directness of introducing the reason for the call after only a brief opening.

However, some comments displayed less skepticism. One of the criticisms that Google received about Duplex was that the digital assistant never self-identified or said it would record the call. Commenters picked up on this, writing, 'I love how Google said they would let people know it's a robot so the critics would shut up yet they made it so subtle at the beginning.' Others were eager about what would come: 'In a year it's gonna be twice as good and twice as good as that in another!! I can't wait.'

Not everyone was convinced that a digital assistant made the call; this example of brims with sarcasm: 'Wow a human speaking to a human, technology these days, what will they think of next. I'm going to coin the term "phone call", yes that's what we'll call this, catchy.' However, such displays of disbelief would not stand unchallenged: 'That you thought it was actually 'a human speaking to a human' shows how far developed the AI and digital voice technology has actually become, and you are far from the only one who has gotten fooled by the voice, mistaking it for a human.'

Through these comments and retorts made about the recorded exchange, we can obtain insights into what happens when technologies begin to mimic human mannerisms. Here, layer upon layer of suspicion and mistrust colored the contrasting interpretations of the caller. In the end, the caller's true nature was uncovered by technology journalists.

To train their AI, Google would also use human callers to create a baseline for the system. It was subsequently disclosed that the caller of this contested exchange was a human. In one sense, we could say that this individual, calling on behalf of someone else to make a dinner reservation, failed the Turing Test. Still, we need to remember that there was nothing particularly damning about the caller's performance. As a call to make a dinner reservation, it was successful without any notable missteps. The produced understanding as an achievement of the interaction was well on par with the example of the haircut appointment.

³ https://www.youtube.com/watch?v=x_FuvIwSxT4

Rather, it is on the level of classification that the caller is understood to be artificial. With Google's announcement that it would use AI for the service, a seed of suspicion was planted. With this suspicion, even ordinary conversational parties are now classified as artificial agents. It thus seems that knowledge about the *prospect of deceit* is consequential for the interpretation of the actions, well beyond what can be gleaned through the turns at talk. With knowledge of Google's capabilities, new aspects are introduced as potentialities.

Even so, evidence supporting the caller-as-a-bot classification could *subsequently* be garnered from the details of the interaction. The relatively long pauses preceding the utterances of the caller could be heard as one of the many unremarkable details that differentiate one call from the next; however, in online comments, they are presented as evidence of the caller being a bot. When heard as a bot, what is described as directness becomes motivated by the social ineptness of a technical system. Alternatively, when the same sequence is heard as said by a human, it is simply directness. In line 15, there is a self-repair: the beginning of the response 'it's' to the question 'what's your name', is changed to 'Her first name is Anna.' When approached with suspicion, this utterance could be interpreted as a bug in the dialogue system, or, perhaps, as clever way of hiding the fact that the caller is a bot by incorporating the mistakes and self-repair mechanisms of humans in the production of the talk.

To some extent, these alternate hearings of the party talking 'as human' versus 'as bot' could be discussed in terms of aspect perception. Wittgenstein (1953) used the famous example of the duck-rabbit to distinguish between the continuous seeing of a single aspect and the dawning realization of when a second aspect becomes available. The continuous seeing of the figure as a picture-rabbit is typically devoid of the aspect sense. Reports on this perception would simply express 'seeing the rabbit'. In contrast, Wittgenstein argued that to say that one now sees the figure *as* a rabbit is to allude to some form of change. 'The expression of a change of aspect is the expression of a new perception and at the same time of the perception's being unchanged' (Wittgenstein, 1953, p. 196). In his view, the complication is not to be resolved by causal inference, since the operation at work is conceptual. Here, the discussion on aspect perception can be coupled with the previously covered argumentation around categorial analysis. A report about the change of aspect seeing from picture-rabbit to picture-duck is another case of the practice of identity redocumentation in the sense that the operation is one of reclassification.

As we move to our final case, we will continue to address the phenomenon of reclassifications in telephone conversations and topicalize changes in aspect hearings from humans to bots.

5.3 Hello, this Is Lenny

Our third example of the variations of contested identities, tied in with the distinction of 'caller' and 'called,' will shift the focus from a single conversation to a

Suspicious Minds: the Problem of Trust and Conversational Agents

collection. The corpus is based on a chatbot specifically designed to handle unsolicited telemarketing or scam calls. Unlike the Duplex system, the chatbot Lenny is a computer program that only plays a set of pre-recorded messages to deal with spammers. Despite the apparent simplicity of the design, it has proved very efficient in keeping callers engaged for minutes on end, effectively using up their time.

Marc Relieu and colleagues headed a series of investigations into the interactional operations of Lenny. Among hundreds of recordings made available online, Sahin et al. (2017) selected and transcribed 200 calls for further analysis. These calls averaged 10 minutes of conversation, and the authors sought to understand what made the system so effective. Lenny opens with a greeting: ‘Hello. This is Lenny,’ and the system then listens for a response. When it registers a silence of more than one and a half seconds, it plays a sequence of phrases. The script always continues with the following three expressions, played in turn: ‘Err...sorry, I can barely hear you there,’ ‘Yes, yes, yes,’ and ‘OH good yes yes yes.’ After this, 12 additional phrases of varying lengths are played in a loop until the caller hangs up. So, how can such a small set of phrases put on repeat keep callers on the line for sometimes up to an hour?

Relieu et al. (2019) argued that one key to understanding this lies in the design of the introduction of the calls. Lenny’s second turn employs a practice for dealing with problems or troubles in speaking, hearing, or understanding talk, which has been subsumed under the notion of ‘conversational repairs’ (Schegloff et al., 1977). The specific repair initiation in use here topicalizes hearing troubles, and it appears to tie back to the caller’s last turn. What ensues typically involves efforts to correct the alleged sound quality problems, and Lenny then provides two turns positioned as responses to whatever conversation is produced. According to Relieu et al. (2020), in these initial struggles and their resolutions, Lenny’s humanity is placed beyond all doubt. The sequential understanding on display here effectively establishes Lenny as an older man with a potential hearing disorder.

As pointed out by Sahin et al. (2017), since the early days of ELIZA (Weizenbaum, 1966) and onwards, bots have regularly been ‘built as personas (an artificial but realistic identity) who produce a recognizable type of conduct from the members of such categories (e.g., an “old guy”)’ (Sahin et al., 2017, p. 321). By carefully crafting the features of the bots’ talk, the conversational partner is invited to recognize or infer additional aspects that typically belong to members of the identified category.

As the conversations progress, the troubles with Lenny as an interactional partner gradually pile up. Sahin et al. (2017) reported that in only 5% of their examined calls did the caller explicitly state some realization of talking to an automated system. Regardless, all calls were eventually ended by the caller, and we took an interest in some of the ways this occurred.

Since Lenny used a limited number of pre-recorded phrases, repetition became an issue when the caller hung on long enough. Depending on how much they talked, Lenny reused his turns after about 8–12 minutes. In the following

example,⁴ the caller trying to book a service for an air-conditioning system has been on the phone with Lenny for a little less than 10 minutes. The extract begins with the caller requesting that her recipient give his first name for the eighth time.

Extract 3. Conversation with Lenny

01 CALLER: And how do you spell your first name?
02 (1.9)
03 LENNY: I'm sorry I- I- couldn't quite catch you there=
04 CALLER: =You don't know your name?
05 LENNY: What was that again?
06 (1.0)
07 CALLER: You don't know your na↑me
08 (1.6)
09 LENNY: Eh but eh- sorry a- again
10 (1.7)
11 CALLER: You're going to miss out on saving a lot of money↑
12 LENNY: Could you say that again (.) [again please
13 CALLER: [Is there somebody else
14 there?
15 Is there somebody else there that I can speak with?
16 (1.3)
17 LENNY: Yes yes yes
18 (1.1)
19 CALLER: Yeah because you're no:t- (0.5) you're not being
20 cooperative and this might be something you guys
21 need to get done
22 (1.5)
23 LENNY: Sorry eh which company did you say you were calling
24 from again?
25 CALLER: I already told you that can I speak with somebody
26 else in your household?

In the first few turns of Extract 3, Lenny consecutively deploys several conversational repairs (lines 3, 5, 9, 12), with the caller exhibiting a frustrated tone of voice in return. As she asks for the availability of a different family member, Lenny responds with an emphatic 'Yes, yes, yes' (line 17). The caller's response (line 19) treats Lenny's confirmation as referring to her request and thus meaningfully contributing to the talk. In acknowledging what appears like a moment of shared understanding, she elaborates on his lack of cooperation. The illusion of being on the same page is soon shattered.

⁴ The entire recording is available at <https://youtu.be/Ue8l8yI-vAY>

Suspicious Minds: the Problem of Trust and Conversational Agents

After several shorter turns, Lenny launches one of his more extensive turns (46 seconds long) about the last time someone called and how he got into trouble with his daughter (see Extract 4). This is the second time in this conversation that the same story is being told. There is little turn-taking to speak of here. Still, in the multiple attempts by the caller to break into the conversation, we observe a gradual determination to exit the interaction until she finally gives up and terminates the call.

Extract 4. Continuation from extract 3

26 LENNY: Well [you know] [here's] here's the thing
27 CALLER: [I'm not gonna leave] [sir-]
28 LENNY: ['cause] the last time that I that [someone called]
29 CALLER: [I-] [there is something]
30 LENNY: [up e::r] and spoke to me [on] the [phone]
31 CALLER: [I don't wanna be rude] [oh:] [my:]
32 LENNY: [I got in quite] a bit of trouble from the people here
33 CALLER: [go:d]
34 LENNY: because I [went for something] [that I shouldn't have]
35 CALLER: [you told me that] [you're repeating]
36 LENNY: [e::r I probably shouldn't be] telling you that
37 CALLER: [yourself]
38 LENNY: but e:rm [yes I- I think that] my- my eldest Rachel
39 CALLER: [you already told me]
40 LENNY: [she e:r] e:r [wouldn't] speak to me for a
41 CALLER: [are you listening?] [okay]
42 LENNY: week. now you know [that] [happens you know but] e:r
43 CALLER: [okay] [you already told me that]
44 LENNY: that really hurt an- an- [and sometimes in families]
45 CALLER: [you already told me that]
46 LENNY: these- these things are quite important
47 [you know more important] than any you know
48 CALLER: [o::h my god]
49 LENNY: job [or phone call or] what [or what]ever it is.
50 CALLER: [have a good day sir] [have a good day]
51 (2.1)
52 CALLER: Have a good day. Because you don't wanna do this
53 (1.2)
54 LENNY: Well you know, since [you since you put it that way]
55 CALLER: [I- I- I gotta go]
56 LENNY: [you have been quite friendly] and straightforward
57 CALLER: [I'm gonna let him go, he's retarded]
58 (call ends)

In this example, there are several references to Lenny's propensity to repeat himself: 'you're repeating yourself' (line 35) and 'you already told me' (lines 35, 39, 43, 45). Earlier in this call and in other conversations, the callers would also show that they knew about his family members by now ('yeah, Rachel Larissa whatever'). In coping with the situation, some callers also began orienting to other co-workers in the call center by commenting on what they experienced ('this is the same thing he said,' 'this is exactly the thing he told me when we first talked,' and 'most talkative').

Another practice identified in connection to the conclusions of these calls was the formulation of new categorizations of the type of person the caller was speaking to. These were either addressed directly to Lenny ('you sound like you've got a mental problem') or to a present third party ('I think this guy is senile'). In Extract 4, the caller ends on a similar note: 'I'm gonna let him go, he's retarded' (line 57).

The ultimate re-categorization, however, was of course the realization that Lenny was not a person at all ('This is so much fun. I've never seen anybody have their own routine on the phone. This is quite cool. Since both of us are going to talk *now* I'm thinking maybe this is a recording because you can't hear anything that I'm saying at this point'). These discoveries seemed to be occasioned by Lenny's blatant disregard for monitoring overlapping talk, or when he responded with indexical expressions to sounds and background noise.

As soon as these discoveries were made, it became evident that it was no longer meaningful to address Lenny in the same way as before. The new categorization of Lenny as an automated system suddenly rendered the purpose of the conversation senseless. No longer was there any understanding between the parties. One recourse was to voice the revelations to co-present parties ('This is a recording' and 'I thought it was a real person at first') A caller who suffered through 18 minutes of talk exclaimed, 'This is a recording. I've been talking five minutes to a recording,' clearly understating the amount of time wasted.

A few callers continued the interaction with Lenny after their first realization that he might not be a human. 10 minutes into one of the calls, and after Lenny had repeated the same utterance about his eldest daughter a second time, a caller who had been trying to sell a service asked, 'Are you a robot?' After not getting an answer to his question, the caller repeats the question a second time. From a position of suspicion and curiosity, he continues to interact with the Lenny for more than 20 minutes. Instead of attempting to sell a service as he initially did, however, the caller now tests Lenny by repeating the same phrase and question (e.g., 'How old is your system?', 'How old is your AC system?', 'What's the age of your system?' or 'To top it off with two pounds of corn', 'Two pounds of corn', 'They topped it off with up to two pounds of corn'), noting some of his discoveries ('You don't have any buttons for that one, huh?') and inviting a co-present party to take over the phone ('Here, talk to him. Talk to him for 30 seconds.').

6 Discussion

As described in our examples, conversational systems are typically designed to mimic features found in human communication. They do so to varying degrees, but minimally, there are attempts to establish or hint at some relation between consecutive turns. The interpretative work in such interactions is therefore double, as both sides concurrently analyze ongoing events. The outcome of this process is a collaborative achievement in which individual contributions become, if not meaningless, at least challenging to distinguish.

6.1 Making Sense of Programmed Responses

One early experiment exploring how programmed responses would be dealt with was carried out without computers. Garfinkel (1967) recruited 10 graduate students under the pretense that they would receive counseling:

Each subject was seen individually by an experimenter who was falsely represented as a student counselor in training. The subject was asked to first discuss the background to some serious problem on which he would like advice, and then to address to the “counselor” a series of questions, each of which would permit a “yes” or “no” answer. (Garfinkel, 1967, p. 79)

The principal deception embedded in the experiment was that the sequence of answers had been determined in advance from a table of random numbers. Garfinkel sought to examine how subjects would interpret and make sense of the ‘advice’ given. Among the many observations made, the importance of the sequential positioning of the pre-determined responses was a central finding. The subjects typically treated the experimenter’s answers as answers to the questions. The answers were thus heard as motivated by the questions.

Occasionally, these random responses would be perceived as inappropriate or contradictory, which made the subjects wary. Suspicions would recede if subsequent advice happened to be congruent with the subjects’ positions on the matter to resolve contradictions. If doubt persisted, responses from the counselor lost their character as answers and were instead transformed into events of ‘mere speech’. Garfinkel also noted that ‘those who became suspicious, simultaneously, though temporarily, withdrew their willingness to continue’. (Garfinkel, 1967, p. 91).

From these findings, we can conclude that the students made whatever sense they could out of the situation as unknowing research subjects in an experiment. We surmise that Garfinkel seemed to argue that the ‘yes’ and ‘no’ replies became meaningful because (a) they were heard as responses

to questions posed (a sequential analysis) and (b) they were perceived as grounded in the expertise and experiences of ‘a counselor in training’ (a categorial analysis). Both analyses should be understood as achieved by the students there and then. Despite its age, this experiment is very illuminating for our present-day situation with chatbots and digital agents.

6.2 The Understanding of Relational Agents

Throughout this essay, we have posited a series of queries about our relations to interactive technologies: How can we understand concerted actions in a world of fake actors? Do our interactions with the artificial constitute real or mere simula-cra of understanding? These are fundamental questions that demand a thorough, reasoned account. Our position is that the answers are to be found by studying the practical reasoning of society members’ encounters with these issues as matters of their everyday lives. Members’ practical actions are ‘contingent ongoing accomplishments of organized artful practices’ (Garfinkel, 1967, p. 11) and make up the central object of our investigation.

This investigation, informed by empirical cases and prior studies, shows how situated analyses and categorizations are reflexively related in interactions with other parties. The ascribed forms of understanding that travel with these practices therefore paint a fractured picture. This point can be further illuminated by the therapy chatbot Woebot, described as a ‘relational agent for mental health’ with no deception involved in the packaging. In a New York Times article on the topic, Karen Brown asks, ‘When your therapist is a bot, you can reach it at 2 a.m. But will it really understand your problems?’ (Brown, 2021). It is not argued here that the bot’s performance or that the models for diagnosing mental health issues are flawed. Although the article points out that ‘some automated conversations can be clunky and frustrating when the bot fails to pick up on the user’s exact meaning,’ the critique further emphasizes that effective treatment requires a ‘human-to-human’ connection that automated solutions are unable to provide. However, the findings of an observational study including 36,000 Woebot users ran contrary to this commonsense notion, indicating that the system was surprisingly effective at establishing therapeutic bonds with its users. As the authors argued, ‘study participants reported that they felt cared for by the CA [conversational agent] (e.g., “Woebot felt like a real person that showed concern”), despite the fact that the tool’s scripts reminded users that Woebot is not a real person’ (Darcy et al., 2021). We argue that at the heart of the paradox lies a separation between understanding as dispositional statements versus as a constitutive feature of action. As Sherry Turkle pointed out in the final paragraph of the aforementioned New York Times article, ‘You have created a bond with something that doesn’t know it is bonding with you. It doesn’t understand a thing.’ (Brown, 2021). So, on the one hand, these chatbots can be said to understand language and interact with people; on the other hand, they do not understand a thing.

Suspicious Minds: the Problem of Trust and Conversational Agents

6.3 Trust and Suspicion

In continuing our argument to address the notions of trust and suspicion, we seek to heed the warning proposed by Porcheron et al. (2018) not to anthropomorphize conversational technologies. There is no point in imparting the analytic registers of HCI research with conceptual ambiguities by building technical terminology out of vernacular expressions. As argued by Button et al. (2022), drawing on the work of Winch (1958), such borrowing of natural-language expressions risks creating confusion, since the commonplace uses of these words continue to remain in play.

In the context of much conversational AI use, there is a risk of conceptual conflation where researchers and developers may mischaracterize interactions with the interface *as* conversations. Nevertheless, this reasoning is still premised on the condition that users know that they are interacting with a piece of technology. In exchanges wherein the other party's identity remains unknown, the situation still begs the question of what the interactants *themselves* make of it. Artificial agents can be treated as genuine interactional partners, and, inversely, humans can be downgraded to the status of machines. Which of these options play out in situated instances of use? These are empirical matters to be examined in their occasioned particulars, as outcomes are not given only from the designed features of conversational systems. That is not to say, however, that structural features may not influence moments of use.

In the case of Woebot, the system was designed to adopt a strategy of transparency and to self-identify as a bot. Conversely, systems that build on the opposite approach muddle the waters and place the burden of identification on the other party. This work must build on contextual resources and whatever additional details are made available within the interaction. In the case of voice-based systems, the quality and character of the voice are the primary means of positioning the agent as a person. What this implies is that, based on the voice alone, a battery of social identities can be inferred—ethnicity, age, gender, and class—which are all possible to perceive as aspects when hearing some speech; together, they may inform the classification of the individual as someone belonging to a particular group. Additionally, Pradhan and Lazar (2021) argue that inviting such classifications through the design of conversational agents, as if they had distinct personalities, may also reinforce social stereotypes.

In the illustrations, we outlined one type of situation in which previous events roused misgivings concerning a caller's identity. After Google had publicized its capabilities, later interactions were treated with suspicion from the outset. We also featured the ongoing accomplishment of suspicion as an interactionally occasioned phenomenon. In the first case, issues concerning identity and motives were called into question proactively, and evidence became a matter of justifying the hearing. Still, we would relate this hearing to aspect perception. The reason for not regarding this listening as continuous hearing is that the listener remains aware of the voice's dual nature: 'What sounds like a bot to me, here and now,

is still possible to be heard as a human.’ In contrast, when encountering an evidently synthetic voice, suspicion is futile.⁵

In the Lenny example, suspicion was more reactive. Interactional troubles, incoherencies, or failures to demonstrate understanding as a constitutive feature of the turn-taking organization could prompt identity redocumentation. Uncertainties and ambiguities would thus necessitate the search for new categorizations that could transform bewilderments into expected events. It is generally preferred not to inform recipients of what they already know (Sacks and Schegloff, 1979). While repetitions (such that cannot be reasonably linked to conversational repairs) clearly violate this norm, speakers can be excused if they are categorized as senile. For the callers in the Lenny case, such transformations would likely push the prospect of a possible sell toward the less likely end. In the few instances involving total re-categorizations, from human to technical systems, the operation became a gestalt shift that altered the significance of past events and rendered any continuation of the original enterprise meaningless. In both types of situations, suspicion is characterized by aspect perception—as entertaining the possibility of an alternative.

We could also conceive of trust as the inverse of these practices. In perceptual terms, trust would be characterized by continuous hearing/seeing and as constituting a mode of action where categorizations and identifications remain unproblematic and stable. The idea of trust as a constitutive order—as a necessary background condition *of* and not emerging *from*—action is at the core of the notion, as described by Garfinkel (1963). If people must decide in each instance whether or not to invest trust, this would be a ‘recipe for inaction’ (Watson, 2015), and it is perhaps another way of saying that *we can’t build our dreams on suspicious minds*.

7 Conclusion

One of the key points of this study has been to expand on prior work in the field by focusing on examples with delusive elements. As argued by Reeves and Porcheron (2022), these instances could be considered extreme cases that build the semblance of participation by duping the recipients through various verbal features and the strategic use of disfluencies. We have illustrated how a particular interactional event, such as a minimal response token with a rising intonation, becomes analyzed by overhearing parties as something remarkable—either as a laudable sign of technological progress or as an ethical low-water mark in the exploitation of unknowing recipients.

Another point raised here is the reminder that specific performances or actions do not equate to demonstrations of understanding. In the context of computers, this reasoning has already been eloquently elaborated upon by Button et al.

⁵ That is, unless there would be some reason to suspect that the human is masquerading as a bot.

Suspicious Minds: the Problem of Trust and Conversational Agents

(1995). Whereas the outcomes of computations can certainly mimic human performance, the inferences that we make regarding the understanding of those agents are not necessarily warranted. We wish to stress the renewed relevance of this argument. As the models supporting conversational agents are growing ever more powerful, we increasingly risk being duped if our awareness of these capabilities is not raised in turn.

In addition to questioning how understanding may be demonstrated, we have also introduced a distinction between the two uses of the notion. We have argued that *interactional* use is related to but not identical to *dispositional* use. This separation does not speak of merely theoretical constructs awaiting their application to data extracts. We argue that they are in fact integral to the machinery of talk-in-interaction; more specifically, they are constitutive of the sequential and categorial analyses that are performed by members of a conversation.

Finally, this observation takes us to our last point, which covers the implications for the broader body of conversational AI work in HCI and CSCW, and particularly the ethical implications of conversational agent design in relation to trust and deception. Our first recommendation is to design artificial agents that always self-identify as bots and, if relevant, declare their limitations regarding such human qualities as personal experiences, emotions, or consciousness. If voice is the only medium, it can be manufactured to signal its synthetic nature—it has been demonstrated that such non-human voices can be beneficial for naïve users (Moore, 2017a, 2017b). If this does not happen, and conversational technologies with obscured identities become prevalent, we envisage the following scenario. One possibility is to continue approaching new interactional partners naïvely (i.e., to routinely treat appearances at face value) (Sacks, 1972) and accept all voices as persons until proven wrong. By walking this path, we will inevitably be deceived from time to time. If deceptions are noticed, this may impact our willingness to maintain the attitude. An opposite strategy would be to approach any hearing with suspicion. The problem with this tactic is that an attitude insisting on aspect hearing will likely find the evidence it is looking for—regardless of what is true. This implies that real persons will be mischaracterized as artificial due to our suspicions. Should we accept this as collateral damage? We would disagree. Trading in people’s humanity for the sake of a convenient technological interface seems like a steep price to pay. Consider the judges in the Loebner Prize Contest in Artificial Intelligence who behaved overtly suspiciously, even hostilely, in their interactions (Proudfoot, 2011). While such behavior might be acceptable in a competition, it would stand as a terrible model for social interaction. Yet, if deception, leveraged by technology, is propagated on a massive scale, the ensuing response will likely come in the form of suspicion.

Our second recommendation is directed at professional analysts, whether they be researchers or designers, of conversational agents. Participants in a conversation are already involved in ongoing analysis. They are the first conversational

analysts on the scene, and it is their understanding of the conversation that should be our primary concern. In this work, we can identify the ordinary actions that organize the activity. In line with Button et al. (2022), we are wary of any analysis that substitutes those actions with theoretical and methodological constructs. Secondary analyses are always possible, and as we have demonstrated, they could also be transformed into materials for investigation. However, we should not make the mistake of treating transcripts or recorded data as if they were constitutive of the situations that are studied. Such treatments may overreach their analytic claims and skew our understanding of how conversational agents are used.

Acknowledgments

This work was supported by The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society.

Funding Open access funding provided by University of Gothenburg.

Declarations

Competing Financial Interests The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abedin, Babak; Christian Meske; Iris Junglas; Fethi Rabhi; and Hamid R. Motahari-Nezhad (2022). Designing and Managing Human-AI Interactions. *Information Systems Frontiers*, vol. 24, no. 3, pp. 691–697. <https://doi.org/10.1007/s10796-022-10313-1>
- Brown, Karen (2021, June 1). Something bothering you? Tell it to Woebot. *New York Times*. <https://www.nytimes.com/2021/06/01/health/artificial-intelligence-therapy-woebot.html>
- Button, Graham; Jeff Coulter; John R E Lee; and Wes Sharrock (1995). *Computers, minds, and conduct*. Cambridge, UK: Polity Press.

Suspicious Minds: the Problem of Trust and Conversational Agents

- Button, Graham, Michael Lynch, and Wes Sharrock (2022). *Ethnomethodology, Conversation Analysis and Constructive Analysis. On Formal Structures of Practical Action*. London: Routledge.
- Cambre, Julia; and Chinmay Kulkarni (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on human-computer interaction*, vol. 3, issue CSCW, article no. 223, pp. 1–19. <https://doi.org/10.1145/3359325>
- Chen, Jessie Y. C. (2022). Transparent Human–Agent Communications. *International Journal of Human–Computer Interaction*, vol. 38, no. 18–20, pp. 1737–1738. <https://doi.org/10.1080/10447318.2022.2120173>
- da Silva, Thiago H. O.; Vasco Furtado, Elizabeth Furtado, Marilia Mendes, Virgilio Almeida, and Lanna Sales (2022). How Do Illiterate People Interact with an Intelligent Voice Assistant? *International Journal of Human–Computer Interaction*, pp. 1–19. <https://doi.org/10.1080/10447318.2022.2121219>
- Darcy, Alison; Jade Daniels; David Salinger; Paul Wicks; and Athena Robinson (2021). Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Formative Research*, vol. 5, no 5, pp. e27868. <https://doi.org/10.2196/27868>
- Fischer, Joel E, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland (2019, August 22-23). Progressivity for voice interface design. *Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin*. Article no. 26, <https://doi.org/10.1145/3342775.3342788>
- Garfinkel, Harold (1963). A Conception of, and Experiments with, “Trust” as a Condition of Stable Concerted Actions. In O. J. Harvey (Ed.), *Motivation and Social Interaction. Cognitive Determinants* (pp. 187–238). New York: Ronald Press Company.
- Garfinkel, Harold (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Google Duplex: AI assistant makes a restaurant reservation* (2018). Retrieved 29 March 2021 from <https://kzbin.info/www/google-duplex-ai-assistant-makes-a-restaurant-reservation/rPcppqLrLiripY>
- Jefferson, Gail (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation Analysis. Studies from the first generation* (pp. 13–31). Amsterdam: John Benjamins.
- Landay, James A., Nuria Oliver, and Junehwa Song (2019). Conversational User Interfaces and Interactions. *IEEE Pervasive Computing*, vol. 18, no. 2, pp. 8–9. <https://doi.org/10.1109/mprv.2019.2921176>
- Mair, Michael; and Wes Sharrock (2021). Action, meaning and understanding. Seeing sociologically with Harvey Sacks. In R. J. Smith, R. Fitzgerald, and W. Housley (Eds.), *On Sacks. Methodology, materials, and inspirations*. London: Routledge.
- McDermott, Drew (2007). Level-headed. *Artificial Intelligence*, vol. 171, no. 18, pp.1183–1186. <https://doi.org/10.1016/j.artint.2007.10.013>
- Moerman, Michael; and Harvey Sacks (1988). On “understanding” in the analysis of natural conversation. In M. Moerman (Ed.), *Talking culture: ethnography and conversation analysis* (pp. 180-186). Philadelphia, PA: University of Pennsylvania Press.
- Moore, Roger K. (2017a). Appropriate voices for artefacts: some key insights. *1st International workshop on vocal interactivity in-and-between humans, animals and robots, Skövde, Sweden*. Article no. 8, http://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_8.pdf
- Moore, Roger K. (2017b). Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In K. Jokinen and G. Wilcock (Eds.), *Dialogues with Social Robots. Lecture Notes in Electrical Engineering*, vol. 427, pp. 281–292. Singapore: Springer. https://doi.org/10.1007/978-981-10-2585-3_22

- Porcheron, Martin; Joel E. Fischer; Stuart Reeves; and Sarah Sharples (2018, April 21–26). Voice interfaces in everyday life. *Proceedings of the 2018 CHI conference on human factors in computing systems, Montreal, QC, Canada*. Paper no. 640, pp. 1–12, <https://doi.org/10.1145/3173574.3174214>
- Pradhan, Alisha; and Amanda Lazar (2021). Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. *CUI 2021: 3rd Conference on Conversational User Interfaces, Bilbao (online)*. Article no. 12, pp. 1–4, <https://doi.org/10.1145/3469595.3469607>
- Proudfoot, Diane (2011). Anthropomorphism and AI: Turing’s much misunderstood imitation game. *Artificial Intelligence*, vol. 175, no. 5–6, pp. 950–957. <https://doi.org/10.1016/j.artint.2011.01.006>
- Randall, David; Mark Rouncefield; and Peter Tolmie (2021). Ethnography, CSCW and Ethnomethodology. *Computer Supported Cooperative Work (CSCW)*, vol. 30, no. 2, pp. 189–214. <https://doi.org/10.1007/s10606-020-0938>
- Reeves, Stuart; and Jordan Beck (2019). Talking about interaction*. *International Journal of Human-Computer Studies*, vol. 131, pp. 144–151. <https://doi.org/10.1016/j.ijhcs.2019.05.010>
- Reeves, Stuart; and Martin Porcheron (2022). Conversational AI: Respecifying participation as regulation. In W. Housely; A. Edwards; R. Benetio-Montagut; and R. Fitzgerald (Eds.), *Handbook of Digital Society*. (pp. 573–592) London: SAGE Publications. <https://doi.org/10.4135/9781529783193>
- Relieu, Marc; Merve Sahin; and Aurelien Francillon (2019). Lenny the bot as a resource for sequential analysis: exploring the treatment of Next Turn Repair Initiation in the beginnings of unsolicited calls. *Proceedings of the Mensch und Computer 2019, Hamburg*. <https://doi.org/10.18420/muc2019-ws-645>
- Relieu, Marc, Merve Sahin, and Aurélien Francillon (2020). Une approche configurationnelle des leurres conversationnels [A configurational approach to conversational lures]. *Réseaux*, vol. 220–221, no. 2, pp. 81–111. <https://doi.org/10.3917/res.220.0081>
- Ryle, Gilbert (1949). *The concept of mind*. London: Hutchinson’s University Library
- Sacks, Harvey (1972). Notes on police assessment of moral character. In D. Sudnow (Ed.), *Studies in Social Interaction* (pp. 280–293). New York: The Free Press.
- Sacks, Harvey (1992a). *Lectures on conversation, volume 1*. Cambridge, MA: Blackwell.
- Sacks, Harvey (1992b). *Lectures on conversation, volume 2*. Cambridge, MA: Blackwell.
- Sacks, Harvey, and Emanuel A Schegloff (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 15–21). New York: Irvington.
- Sahin, Merve, Marc Relieu, and Aurelien Francillon (2017). Using chatbots against voice spam: Analyzing Lenny’s effectiveness. *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017), Santa Clara, CA*. pp. 319–337, <https://www.usenix.org/system/files/conference/soups2017/soups2017-sahin.pdf>
- Schegloff, Emanuel A. (1982). Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University roundtable on languages and linguistics* (pp. 71–93). Washington, DC: Georgetown University Press.
- Schegloff, Emanuel A. (2007). *Sequence organization in interaction. A primer in conversation analysis*. Cambridge, MA: Cambridge University Press.
- Schegloff, Emanuel A, Gail Jefferson, and Harvey Sacks (1977). The preference for self-correction in the organisation of repair in conversation. *Language*, vol. 53, pp. 361–382.
- Schutz, Alfred (1976). *Collected papers II. Studies in Social Theory*. Hague: Martinus Nijhoff.
- Stahl, Gerry (2016). From intersubjectivity to group cognition. *Computer Supported Cooperative Work (CSCW)*, vol. 25, nos. (4–5), pp. 355–384. <https://doi.org/10.1007/s10606-016-9243-z>

Suspicious Minds: the Problem of Trust and Conversational Agents

- Turing, Alan M. (1950). Computing Machinery and Intelligence. *Mind*, vol. 54, no. 236, pp. 433–460.
- Watson, Rod (1994). Review article: Harvey Sacks's sociology of mind in action. *Theory, Culture and Society*, vol. 11, no. 4., pp. 169–186.
- Watson, Rod (2009). Constitutive practices and Garfinkel's notion of trust: Revisited. *Journal of Classical Sociology*, vol. 9, no 4, pp. 475–499. <https://doi.org/10.1177/1468795X09344453>
- Watson, Rod (2015). De-reifying categories. In R. Fitzgerald and W. Housley (Eds.), *Advances in Membership Categorisation Analysis* (pp. 23–50). London: Sage.
- Weizenbaum, Joseph (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, vol. 9, no 1, pp. 36–45. <https://doi.org/10.1145/365153.365168>
- Wiggers, Kyle (2018). Google's Duplex is rolling out to Pixel owners – here's how it works. *Venture Beat*. <https://venturebeat.com/2018/11/21/googles-duplex-is-rolling-out-to-pixel-owners-heres-how-it-works/>
- Winch, Peter (1958). *The Idea of a Social Science and Its Relation to Philosophy*. London: Routledge and Kegan Paul.
- Wittgenstein, Ludwig (1953). *Philosophical investigations*. Oxford: Blackwell.
- Wowk, Maria T., and Andrew P. Carlin (2004). Depicting a Liminal Position in Ethnomethodology, Conversation Analysis and Membership Categorization Analysis: The Work of Rod Watson. *Human Studies*, vol. 27, no. 1, pp. 69–89. <https://doi.org/10.1023/B:HUMA.0000012245.95451.4b>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.