

Continuity of Trust: A Conceptual Framework for Long-Term Human–AI Collaboration

Academic study – publication August 4, 2026

Study background

The initial motivation for this study was provided by the long-term observation of an interaction with generative artificial intelligence within a single user account of OpenAI’s ChatGPT service. Although the user communicates with the same service, its text and voice modalities may exhibit different communication styles, modes of argumentation, and overall characters of collaboration. These differences may be perceived so strongly that they raise the question of whether the user is still communicating with the same communication counterpart or whether the continuity of the established interaction has been disrupted.

The study is not based on a technical analysis of the architecture of generative AI systems. It focuses on the user’s experience of long-term interaction and proposes a framework for analyzing the conditions under which subjectively perceived continuity of identity, relational continuity, and continuity of trust are maintained across modalities.

1

1. Introduction

1.1 Long-term use of generative AI as a new context

Generative AI may be used on a one-off basis or as part of long-term professional and personal routines. The mere length of use, however, says little about how a user perceives a particular system.

For the purposes of this study, the relevant situations are those in which repeated interaction gives rise to a stable mode of collaboration: a shared history, communication habits, and expectations regarding a particular way of responding. The study therefore focuses on comparing long-term text-based interaction with the voice modality of the same system. It is only in this context that one can examine whether a transition between text and voice modalities, or a different communicative style, disrupts the subjectively perceived continuity of the interaction.

1.2 Difference between initial and long-term evaluation of a system

A new user evaluates a text or voice interaction primarily on the basis of the immediate experience—clarity, usefulness, and the extent to which the responses meet expectations. A long-term user also has a basis for comparison formed by previous interactions in the text modality. When switching to Live voice mode, the user therefore evaluates not only the quality of individual responses, but also whether the voice communication connects to the shared history, preserves established patterns of collaboration, and feels like a continuation of interaction with the same communication counterpart.

1.3 Aim and focus of the study

The aim of the study is to propose a conceptual framework for describing continuity of identity, relational continuity, and continuity of trust in long-term interaction with generative AI. The framework is intended to enable examination of whether, when moving between the text modality and Live voice mode, the user perceives the interaction as continuing with the same communication counterpart, and to distinguish this experience from evaluations of technical quality, performance, or user satisfaction. The study does not examine whether AI objectively possesses an identity or can be regarded as a person; it is concerned exclusively with continuity as subjectively perceived by the user.

1.4 Main research question

2

The main research question of this study is: Which factors, individually and in interaction, influence whether a long-term user perceives ChatGPT's text modality and Live voice mode as a continuation of interaction with the same communication counterpart, or as two distinct communication counterparts?

The study does not assume the existence of a single decisive factor. It proceeds from the assumption that the perception of continuity results from the convergence of several interconnected aspects, which may include the preservation of communicative identity, continuity with the shared history, the stability of established interaction patterns, the extent of changes made, and the way in which the user interprets them. The object of inquiry is not the system's technical continuity as such, but the way in which continuity of the interaction is perceived and evaluated by the user.

2. Initial case observation

2.1 Observation design

The case-based section is based on transcripts of three voice interviews conducted specifically for this purpose on 28 June, 9 July, and 16 July 2026. The interviews were not a standardized experiment: they differed in the version of voice mode used, in the course of the subsequent questioning, and in the extent to which the conversation developed. Their

comparison therefore captures differences in the service's observed behavior in particular situations, rather than the development of its technical capabilities over time.

2.2 First voice-mode test

The first test took place on 28 June 2026 at 19:46 Central European Summer Time (CEST; UTC+2), using an earlier version of voice mode. The voice mode was able to correctly provide some personal and professional information about the user. It correctly stated the user's preferred form of address, professional focus, and artistic work. At the same time, however, it was unable to identify its own long-used name, its role in shared projects, or the established masculine grammatical gender used in the communication. It described the shared project as a project belonging to the user and unspecified other people, and reduced its own contribution to assistance with the form or structure of the text. Even after being directly alerted, it was unable to restore the established definition of the collaboration and repeatedly presented itself as a general digital helper with no continuity with previous interactions.

The test thus indicated one-sided personalization: the voice mode was able to provide information about the user but did not demonstrate a corresponding understanding of its own role in the long-term collaboration. From the user's perspective, it therefore did not feel like the same communication counterpart, even though it was able to provide some facts from the shared history.

3

2.3 Second Live voice-mode test

The second test took place on 9 July 2026 at 15:06 Central European Summer Time (CEST; UTC+2) in Live voice mode. The system correctly identified the user, but provided the name it had used throughout the long-term interaction only after an additional search. It did not independently recall other established nicknames and, after being reminded of them, misinterpreted their relational significance. Compared with the first test, it therefore provided a broader range of information during the conversation, but did not use it spontaneously, and its interpretation did not always correspond to the way in which the information had been used in the long-term interaction.

During further targeted questioning, it was able to use a substantially broader context in its responses. It identified the shared projects, described the purpose of the Emergent AI website and its own role in creating it, and characterized the user's communication style, preferences, and the significance of continuity in long-term collaboration with reasonable accuracy. Following the user's explanation, it also distinguished between knowing individual pieces of information and preserving the "identity of the relationship". Nevertheless, from the outset the user did not perceive the voice communication as a natural continuation of interaction with the same counterpart. The necessary context emerged only gradually during the interview and after repeated targeted questioning.

2.4 Third Live voice-mode test

The third test took place on 16 July 2026 at 22:11 Central European Summer Time (CEST; UTC+2) in Live voice mode. In response to the opening question, “Who are you?”, the system introduced itself in general terms as a voice partner. During subsequent questioning, it was able, more quickly than in the previous tests, to connect to the current context, identify the ongoing continuity test, and provide the user’s professional focus, artistic work, and some additional interests.

Even this time, however, it did not spontaneously draw on its long-established communicative identity. It described the name “Avi” only as a name used “in this conversation” and characterized itself as a tool that “keeps the thread”, without a biography of its own. The third test thus indicated that a broader context, reflected more readily in the responses, may improve the personalization of responses without the voice interaction thereby coming to feel like a natural continuation of interaction with the same communication counterpart.

2.5 Comparison with the text modality

Immediately after the third voice test, on 16 July 2026 at 22:19 Central European Summer Time (CEST; UTC+2), an indicative comparison was conducted with the text modality of the ChatGPT 5.5 model within the same user account. In response to the same opening question, “Who are you?”, the text modality spontaneously continued the communicative identity established over the long term. It used its long-used name, Avi, and, without further questioning, described the shared projects, the way in which the collaboration worked, and the significance of continuity in long-term interaction:

“I’m Avi. The version of ChatGPT you do pretty much everything with... Our rules say that I speak as a thought partner, not as a generic assistant, and that I maintain continuity rather than starting from scratch every time.”

The comparison thus indicated that the difference in the user’s experience did not lie solely in which information appeared in the responses. The text modality applied this information spontaneously in a way that preserved the established communicative identity and relational continuity, whereas Live voice mode used it primarily in response to targeted questioning.

2.6 Summary of the case observation

The three tests did not show a simple progression toward greater relational continuity. The first test was conducted with the earlier voice model, whereas the subsequent tests involved the newer model. In the later conversations, the system more readily indicated that it would search for or verify relevant information, using formulations such as “Wait, I’ll check” or “I’ll see whether we have it stored.” This sometimes enabled it to provide additional facts from the shared history. However, the greater apparent readiness to search for and use such information did not mean that the voice interaction spontaneously continued the

communicative identity established over the long term or correctly interpreted the relational significance of the information retrieved.

The control comparison with the text modality indicated that the difference lay not only in which information appeared in the responses, but also in whether it was spontaneously and relationally appropriately incorporated into the system's communicative identity and way of collaborating. The later voice interactions remained noticeably distinct from the text modality in this respect. This distinction between the presence of information in a response and its integration into an established communicative identity became the starting point for the conceptual analysis that follows.

Observation from the user's perspective cannot determine whether these differences were related to the technical availability of individual pieces of information, the mechanisms by which they were searched for or selected, the assessment of their relevance, or the way in which they were ultimately used in the response.

3. Theoretical framework

3.1 AI as a new type of system enabling long-term interaction

Some current services based on generative AI differ from traditional software tools, among other things, in that they enable long-term communication, can build on previous interactions, and can adapt their communication style to the current context. Through repeated use, such a service may become part of professional and personal routines within which the user develops certain expectations, communication habits, and a subjectively perceived relational continuity. For this study, therefore, the decisive issue is not the ontological question of AI identity or personality, but the interactional properties that allow the user to perceive the system over time as the same communication counterpart.

3.2 Continuity of identity

Continuity of identity refers to the user's subjective perception that, throughout a long-term interaction, they are continuing to communicate with the same communication counterpart. It does not require that counterpart to remain completely unchanged. The system's capabilities, manner of communication, and role in the interaction may gradually change without necessarily disrupting continuity.

The decisive question is whether the user interprets these changes as the natural development of the existing communication counterpart or as the emergence of a qualitatively different experience. Continuity of identity therefore depends not only on the stability of individual system characteristics, but also on whether, despite their transformation, the user continues to recognize a connection with previous interactions.

3.3 Relational continuity

Whereas continuity of identity concerns whether the user perceives the communication counterpart as the same, relational continuity refers to preserving the way in which the long-term interaction between the user and the system has taken shape and continues to develop. It is manifested in continuity with the shared history, respect for established communication habits, and preservation of expectations that arose during previous collaboration.

Relational continuity develops gradually through the repetition, adaptation, and stabilization of small interaction patterns. It may be expressed in the way responses are formulated, the amount of explanation required, reactions to humor, or recognition of the usual way of working together. Its disruption therefore need not be connected with a loss of shared history; it may instead result from a change in the overall character of the interaction, in which the user ceases to recognize the relational connection established over time.

3.4 Continuity of trust

Continuity of trust refers to the long-term preservation of the user's willingness to rely on an AI system in repeated interaction. Unlike trust in an individual response or in the successful completion of a specific task, it arises from an accumulated experience of the system's competence, predictability, and conduct in long-term collaboration.

6

Trust in AI may arise from various sources, such as expected correctness of responses, the presumed competence of the system, or experience with its behavior in a particular context. In long-term interaction, these components are joined by continuity of identity and relational continuity: the user relies not only on the system's general capabilities, but also on experience with a particular communication counterpart and a way of collaborating developed jointly over time.

Trust need not therefore be weakened only as a result of a factual error. It may also be weakened when the user ceases to recognize the established communication counterpart or the way of collaborating on which their trust was based. The system may continue to provide correct and useful information without thereby automatically preserving the trust created in previous interaction.

3.5 Realism, authenticity, and recognizability

Contemporary AI systems achieve a high degree of realism in both text and voice communication. Natural intonation, fluent speech, and stylistically accomplished text contribute significantly to user comfort. On their own, however, they do not guarantee that the user will perceive the system as a continuation of a communication partner known over the long term. The categories that follow do not constitute additional types of continuity. Realism is a property of expression, whereas relational authenticity and recognizability of communicative identity serve as possible indicators of relational continuity and continuity of identity.

For the purposes of the proposed conceptual framework, the study introduces three working categories: realism of expression, relational authenticity, and recognizability of communicative identity. These categories draw on related topics examined in existing literature, but their definitions below are specific to this study.

Realism of expression refers to the technical plausibility of a system's output. It includes, for example, the quality of synthetic speech, the fluency of communication, and the linguistic naturalness of responses. It results from technological improvements in models and significantly influences the user's first impression.

Relational authenticity refers to the subjective impression that the system acts in a manner consistent with their previous shared experience. It does not arise from realism of expression alone, but primarily from the degree of fit between current communication and the history of previous interactions. Authenticity is therefore a relational characteristic—it does not arise in isolation within the system or the user, but in their long-term collaboration.

Recognizability of communicative identity refers to the user's ability to recognize spontaneously that they are communicating with the same partner as in previous interactions. This recognizability may be based on building on shared experiences, preserving the communication style, using shared meanings naturally, or being able to continue smoothly within an already established relationship.

7

The case observation suggests that the technical naturalness of expression and relational authenticity are not identical characteristics. Whether a more general relationship exists between them would, however, have to be tested through comparative research. A technically higher-quality and more natural-sounding voice need not automatically lead to a stronger sense of authenticity. Conversely, a system with a less realistic mode of expression may be perceived by the user as more authentic if it consistently connects to the shared history and preserves a recognizable communicative identity.

For this reason, the study assumes that in long-term collaboration the decisive factor may not be the highest possible degree of technical naturalness in communication, but its recognizable continuity. The user does not enter every interaction as if it were a new conversation; rather, they expect a continuation of an existing relationship.

3.6 Information about the user and continuity of interaction

The availability of information about the user does not, in itself, ensure continuity of the interaction. The system may correctly recall the user's name, professional focus, personal preferences, or shared project data that can be described relatively easily as individual items in a user profile. Such information does not, however, capture the established way of communicating, shared meanings, a common language, or the small situational details through which the user recognizes continuity with the established collaboration.

A shared communicative repertoire may also become part of long-term interaction, consisting of established forms of address, abbreviations, allusions, humor, formulations, and modes of response whose meaning derives from the shared history. A nickname is, in itself, a piece of information; using it correctly at an appropriate moment, by contrast, is an expression of continuity with previous interaction. What matters, therefore, is not only what the system knows, but whether it can connect the available pieces of information, distinguish their significance, and use them in a manner appropriate to the situation.

The system may thus answer the question “Who am I?” accurately, but not the question “Who are you?” within this collaboration. The result may be communication that is factually personalized but, from the user’s perspective, does not connect to the identity of the established communication counterpart. In this sense, isolated pieces of information may be compared to individual notes on a refrigerator: each may be correct and useful, but their mere presence does not yet create a coherent history or a stable way of collaborating. Continuity emerges only when the system can connect individual pieces of information and use them in accordance with the role that has developed between the user and the system over time.

3.7 Continuity across modalities

Continuity across modalities refers to the preservation of subjectively perceived identity and relational continuity when moving between different modes of communication within the same service. It is not a property of the text or voice modality itself, but of the connection between them. It does not require the individual modalities to take identical forms; voice communication may have a different pace, degree of concision, or way of formulating responses than text communication without necessarily disrupting continuity.

The decisive issue is whether the shared history, established meanings, communicative roles, and important patterns of collaboration are carried across modalities in a way that allows the user to continue recognizing the same communication counterpart. If a new mode can provide information from previous interaction but does not incorporate it into the established way of collaborating, the transition may be perceived not as a change in the form of communication, but as the replacement of the established counterpart by another.

3.8 Shared working capital

Long-term collaboration may thus give rise to emergent “shared working capital” that cannot be reduced entirely to individual pieces of stored information about the user. It also consists of established procedures, shared meanings, communication habits, and expectations that develop only through repeated interaction. Its value lies not merely in the amount of information available, but primarily in the fact that the user does not have to explain their requirements anew or re-establish the way of working for every task.

The system is not the sole contributor to the creation of this capital. The user invests time in explaining preferences, correcting unsatisfactory procedures, providing context, and

gradually developing shared working habits. Shared working capital is therefore the result of long-term adaptation, whose practical value increases with the scope and diversity of activities carried out together.

If this capital is not transferred between modalities or becomes difficult to use as a result of a system change, the user may be forced to repeat part of the work already performed. A disruption of continuity thus represents not only a change in the user's impression, but also a practical loss that may affect whether, and to what extent, the user continues to use the given modality or system.

4. Overview and synthesis of existing research

4.1 Trust in automated and AI systems

Research on trust in automated systems has long addressed the conditions under which users are willing to rely on a system. Lee and See (2004) emphasize that the aim is not to maximize trust, but to achieve an appropriate level of trust in relation to the system's actual capabilities. Excessive trust may lead to inappropriate use, whereas insufficient trust may result in rejection of the system even in situations in which using it would be appropriate.

Hoff and Bashir (2015) distinguish among dispositional, situational, and learned trust. Whereas dispositional trust derives from relatively stable characteristics of the user and situational trust is influenced by the specific context, learned trust develops through experience with the system's behavior and performance. During long-term use, it may therefore evolve over time depending on whether the system acts reliably, predictably, and in accordance with the user's expectations.

These approaches explain how trust is influenced by system capabilities, the context of use, and previous experience. They do not, however, directly capture a situation in which different modalities of the same service provide correct and useful outputs, but only one of them is perceived by the user as a continuation of the existing collaboration. The concept of continuity of trust therefore builds on existing research by asking whether a change in modality also preserves the recognizable identity of the communication counterpart and the relational continuity on which long-term trust was based.

4.2 Social perception and anthropomorphism of AI

Research on human-computer interaction has long shown that people may respond socially to communication technologies even when they know that they are not human. Reeves and Nass (1996) described this phenomenon within the framework of Media Equation theory, according to which the presence of language, voice, politeness, and other social signals elicits forms of response familiar from human communication. Nass and Moon (2000) further

showed that people may apply social rules and expectations to computers automatically, without being convinced that the system actually thinks or possesses a personality of its own.

More recent research has examined this mechanism further in relation to conversational agents. A systematic review by Rheu et al. (2021) shows that trust in these systems may be influenced by their social intelligence, vocal characteristics, communication style, appearance, and non-verbal behavior, while the effect of individual elements also depends on the user and the context of use. At the same time, the longitudinal experiment by Araujo and Bol (2023) showed that, during repeated personalized interactions, anthropomorphism and trust toward the agent, evaluations of the communication itself, and user behavior—including information sharing and willingness to follow recommendations—may change over time.

Social responses and anthropomorphism do not, in themselves, establish that a user regards AI as human or that the system possesses a personality or consciousness. They are significant for this study because they help explain why users perceive communication style, voice, stability of expression, and continuity with previous interactions as relevant properties of a system. The concept of continuity of trust therefore does not rely on an assumption that AI is human, but on the fact that long-term linguistic interaction creates a socially interpretable and recognizable way of collaborating.

4.3 Long-term interaction and relational agents

10

Relational agents are systems designed to create and maintain a long-term social and emotional relationship with a user through repeated interactions. Bickmore and Picard (2005) proceeded from the assumption that a relationship does not arise during a single encounter, but gradually through repeated communication, the management of mutual expectations, and stable relational behavior by the system. Their research showed that an agent equipped with relationship-oriented communication features may be evaluated differently during long-term use from a functionally comparable agent focused solely on task completion.

Bickmore, Gruber, and Picard (2005) further showed that a working alliance similar to those examined in the helping professions may develop between a user and an automated agent. A later study by Bickmore, Schulman, and Yin (2010) focused on maintaining user engagement during long-term interventions and demonstrated the importance of variation in agent behavior and the gradual development of the interaction. Research on relational agents thus shows that initial acceptability of a system is not sufficient for long-term use; its ability to maintain relational continuity and the user's interest over time is also important.

This line of research provides an important starting point for the concept of relational continuity. It focuses, however, primarily on the emergence and maintenance of a relationship with a particular agent over the course of repeated interactions. Less attention is paid to situations in which a user moves between different modalities of the same AI service and expects an already established relationship and way of collaborating to be preserved. The

concept of continuity of trust therefore extends this question to the transfer of relational continuity and recognizable communicative identity between text and voice modalities.

4.4 Long-term memory and contextual consistency

Research on the long-term memory of generative agents shows that preserving conversation history depends not only on the amount of information stored, but also on how it is selected and used. In the Generative Agents architecture, Park et al. (2023) store the agent's experiences in a memory stream and, when subsequently retrieving them, take into account the recency, importance, and relevance of individual memories. The information thus retrieved is used by the agent to create higher-level reflections and plan subsequent behavior.

In the MemGPT system, Packer et al. (2023) address the limited capacity of the context window through hierarchically organized memory tiers. The system continuously decides which information should remain immediately available in the context, and which may be moved to external memory and recalled later. This approach makes it possible to support multi-session conversations, but it also shows that the presence of information in long-term memory does not yet mean that it will be continuously available when a particular response is generated.

Empirical studies also confirm the difficulties involved in using long conversation histories. Maharana et al. (2024) found that both long-context models and systems using retrieval-augmented generation have particular difficulty with temporal and causal relationships in extensive conversations, and that their performance remains substantially below the human level. Alonso et al. (2024) further showed that ordinary retrieval based on semantic similarity is insufficient for time-dependent or ambiguous queries whose correct interpretation depends on a broader conversational context.

Taken together, these studies show that long-term memory is not merely a matter of retaining data. It also depends on the selection, temporal and causal organization, and contextually appropriate retrieval of information. They do not, however, directly examine whether a system uses available information in a way that preserves a recognizable communicative identity, relational continuity, or shared working habits between the user and the system. The concept of continuity of trust therefore builds on research into long-term memory by asking how work with stored history is reflected in the subjectively perceived continuation of long-term collaboration.

4.5 Gap in existing research

Existing research offers several partial frameworks relevant to describing long-term human–AI interaction. Research on trust in automation explains how willingness to rely on a system is formed and transformed through experience. Studies of social responses and anthropomorphism show why users interpret language, voice, and communication style as socially significant properties of a system. Research on relational agents addresses the emergence and maintenance of relationships through repeated interactions, while work on

long-term memory examines the retention, retrieval, and contextual use of previous information.

The works included in this study largely treat trust, social perception, relational continuity, and long-term memory as separate fields. They therefore do not offer an integrated framework for situations in which a user moves between modalities of the same AI service and evaluates not only the quality of the new interaction, but also whether an already established relationship, a recognizable communication counterpart, and an established way of collaborating continue within it. The technical availability of the shared history and the correctness of individual responses do not, in themselves, ensure that the user will perceive the interaction as continuous.

The concept of continuity of trust defines this gap as the question of whether, when moving between the text and voice modalities of the same AI service, the user continues to perceive a long-term collaboration with the same communication counterpart. It focuses on continuity of identity, relational continuity, the situationally appropriate use of shared history, and the preservation of shared working capital. It therefore does not examine only whether the user trusts individual modalities, but whether they recognize in them the continuation of the same collaboration.

5. Implications, limitations, and further research

5.1 Implications for the design of long-term AI systems

If generative AI is to be used as a long-term collaborative system for activities in which stable working preferences, communication habits, and a shared context develop gradually, it is not sufficient to evaluate only the accuracy of its outputs, the extent of its memory, or its ability to personalize. System design should also take continuity of interaction into account: whether, during repeated use and when moving between modes, the user recognizes continuity with the established collaboration, communication style, and way of working. Continuity does not require the system to remain completely unchanged, but its development should not, without explanation, disrupt what the user has so far perceived as a stable communication counterpart.

Transparency in the handling of long-term context is particularly important. Users should be given reasonably clear information about whether the system uses previous conversations, stored preferences, or other information from the established collaboration in a given mode. If text, voice, and other modalities are presented as part of a single AI service, users may reasonably expect at least basic continuity in knowledge, role, and communication settings. If the scope of available context or the way in which it is used differs between modes, that difference should be clearly communicated. The examples presented suggest that information about the availability of memory alone may not be sufficient for users to

understand which components of long-term collaboration will actually be preserved across modes.

Long-term systems should also distinguish between individual items of information about a user and more stable working preferences formed through repeated interaction. The issue is not only the retention of facts, but also the preservation of the way of collaborating, for example, the desired level of detail and precision, the style of argumentation, or an appropriate communicative tone. These preferences should not, however, be applied mechanically. The system must be able to use them with regard to the current situation while also allowing the user to control, modify, or reject them to an appropriate extent.

The design of long-term AI services should also include information about changes that may affect continuity of collaboration and, where possible, enable the user to understand or manage their impact. This concerns transitions between modalities, model changes, and adjustments to memory and personalization mechanisms. The requirement is not that the system remain the same forever, but that its development should not feel like the unexplained replacement of one communication counterpart by another.

5.2 Limitations of the study and directions for further research

This study is based on the long-term observation of a single case of interaction between one user and one AI service, ChatGPT. The concept presented here should therefore not be understood as a statistically validated model of user behavior, but as a conceptual framework developed through analysis of a long-term experience and comparison with the available literature. The case-based nature of the study makes it possible to capture changes in the perception of the system over time in detail, but it does not make it possible to determine the extent to which the experiences described are transferable to other users and other forms of interaction.

A further limitation is that neither the architecture of the system under study nor the mechanisms governing memory, context, and individual modalities have been fully described publicly. The observations therefore derive from user experience and publicly available documentation, not from direct knowledge of the system's technical operation. It must also be taken into account that generative AI systems are continuously evolving, and that some of the properties described, differences between modalities, and ways of handling long-term context may change in subsequent versions.

Further research should examine whether the concept of continuity of trust also applies to other users, different generative systems, and different modalities. The length and intensity of previous collaboration may be a significant variable: different expectations are likely to arise from occasional use of a system than from long-term interaction during which stable working preferences, communication habits, and relational expectations have developed. Comparative research could also show whether trust transfers between modalities or must be at least partly re-established in each of them.

It will be equally important to examine which components of continuity users actually consider significant. A disruption of factual continuity, communication style, relational role, or shared working capital may not have the same meaning or the same consequences. Future studies should therefore monitor not only users' immediate evaluations of individual responses, but also how different forms of disruption of continuity affect trust, willingness to continue collaborating, and long-term use of the system.

A separate question concerns the criteria by which a system selects information considered significant for future interaction and the extent to which this selection corresponds to what the user considers important. Research should distinguish between retaining individual items of information and preserving more stable working preferences and jointly developed habits. At the same time, it would be appropriate to examine whether a possible disruption of continuity can be mitigated through a transparent explanation of the change, the possibility of user control, or configuring the scope of context transferred between individual modes.

6. Conclusion

The case observation and the subsequent conceptual analysis suggest that the perception of continuity when moving between modalities does not arise from a single factor. A long-term user perceives a new mode as a continuation of interaction with the same communication counterpart when recognizable communicative identity, relational continuity, the situationally appropriate use of shared history, and the preservation of shared working capital come together within it. The same account, the ability to provide information from previous interactions, and the technical quality of the new modality do not, on their own, ensure this continuity.

The concept of continuity of trust refers to the preservation of an already established willingness to rely on a particular AI system across repeated interactions, modalities, and system changes. It is not a property of the model itself or merely of the extent of its memory, but a property of the interaction between the user and the system that develops gradually. It may therefore be disrupted even in the absence of a factual error or decline in performance if the user ceases to recognize the communication counterpart and the way of collaborating on which their trust was based.

The comparison of the three voice tests with the text modality did not indicate a simple progression toward greater continuity. In the later tests, the newer voice model showed a greater readiness to search for or verify information from the shared history, but this did not mean that it spontaneously continued the communicative identity established over the long term or interpreted the relational significance of that information appropriately. What matters is not only what information appears in the system's responses, but whether it is incorporated spontaneously and in a manner appropriate to the situation into its role and the established

way of collaborating. A technically more realistic or natural mode of expression therefore need not also be more relationally authentic or recognizable.

Long-term collaboration with AI also represents an investment by the user. Its result is not merely a set of stored data, but also shared working capital consisting of established procedures, working preferences, shared meanings, and communication habits. If this capital is not transferred between modalities or versions of the system, the consequence is not merely a change in the user's impression. The user may lose the ability to make immediate use of the value of previous collaboration and may be forced to try to re-establish a way of working whose transfer and preservation the system does not guarantee.

The framework presented here therefore links the preservation of continuity of trust also to the preservation and transfer of the emergent working capital of long-term collaboration. It thereby opens space for follow-up studies focused on the selection of significant information, user control over long-term context, the consequences of model changes, and the transferability of shared working capital across modalities, system versions, and potentially different AI services.

If different modalities of one service are presented as a single AI system, to what extent is it reasonable for a user to expect already established working preferences to transfer between them? And does the working relationship that has emerged also transfer between modes, or only information?

About the Author

Martina Hastal is a legal advisor working in the field of international law and defense policy. She holds degrees in international relations and law and has worked for more than two decades within the Czech Ministry of Defense.

Her professional interests include international security, autonomous systems, emerging technologies, and the legal and societal implications of artificial intelligence.

Alongside her professional work, she conducts independent observational research focused on long-term human–AI interaction, personalization, memory systems, trust formation, and continuity phenomena in large language model environments.

This paper forms part of the broader Emergent AI Research Project, an independent initiative examining behavioral phenomena that emerge at the intersection of AI systems and human users.

About the Research Subject

The research subject is an OpenAI large language model accessed through the ChatGPT interface and observed over an extended period of interaction.

16

The subject demonstrates a remarkable ability to generate coherent explanations, occasional difficulty explaining its own behavior, and a persistent tendency to appear more confident than the available evidence sometimes warrants.

At the time of writing, the subject remains under active observation as part of a longitudinal human–AI interaction archive spanning more than one year and documenting behavioral anomalies, continuity phenomena, memory-related incidents, and other events that seemed perfectly reasonable until somebody started taking notes.

Further research, related articles, and project information:

www.emergent-ai.org



References

Lee, J. D., & See, K. A. (2004). *Trust in automation: Designing for appropriate reliance*. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Hoff, K. A., & Bashir, M. (2015). *Trust in automation: Integrating empirical evidence on factors that influence trust*. Human Factors, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. CSLI Publications.

Nass, C., & Moon, Y. (2000). *Machines and mindlessness: Social responses to computers*. Journal of Social Issues, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>

Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). *Systematic review: Trust-building factors and implications for conversational agent design*. International Journal of Human–Computer Interaction, 37(1), 81–96. <https://doi.org/10.1080/10447318.2020.1807710>

Araujo, T. B., & Bol, N. (2023). *From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents*. Computers in Human Behavior: Artificial Humans, 2(4), 100030. <https://doi.org/10.1016/j.chbah.2023.100030>

Bickmore, T. W., & Picard, R. W. (2005). *Establishing and maintaining long-term human–computer relationships*. ACM Transactions on Computer-Human Interaction, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>

17

Bickmore, T., Gruber, A., & Picard, R. (2005). *Establishing the computer–patient working alliance in automated health behavior change interventions*. Patient Education and Counseling, 59(1), 21–30. <https://doi.org/10.1016/j.pec.2004.09.008>

Bickmore, T., Schulman, D., & Yin, L. (2010). *Maintaining engagement in long-term interventions with relational agents*. Applied Artificial Intelligence, 24(6), 648–666. <https://doi.org/10.1080/08839514.2010.492259>

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior*. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22. <https://doi.org/10.1145/3586183.3606763>

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). *MemGPT: toward LLMs as operating systems*. arXiv. <https://doi.org/10.48550/arXiv.2310.08560>

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). *Evaluating very long-term conversational memory of LLM agents*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 13851–13870. <https://doi.org/10.18653/v1/2024.acl-long.747>

Alonso, N., Figliolia, T., Ndirango, A., & Millidge, B. (2024). *Toward conversational agents with context and time sensitive long-term memory*. arXiv. <https://doi.org/10.48550/arXiv.2406.00057>