Manuscript – accepted at AI & Ethics 29/10/2026

Who Am I Really Talking To? Authenticity, Transparency, and the Case for "AI Ticks" in AI-Mediated Communication

Davide Battisti, Andrea Lavazza, Andreas Komninos

Abstract

The growing integration of artificial intelligence (AI) into everyday messaging systems is transforming the conditions under which interpersonal communication takes place. As AI-mediated communication (AI-MC) becomes increasingly seamless and context-sensitive, it blurs the boundary between human expression and algorithmic assistance. This development raises fundamental questions about potential challenges to authenticity in personal relationships and possible strategies to address them. In this regard, we make three main contributions. First, we offer a technically grounded analysis of emerging AI-MC architectures, distinguishing between external prompting, keyboard-level assistance, and fully integrated in-app co-authoring systems. Second, we develop a normative account of authenticity in AI-MC, understood as a second-person, relationship-sensitive notion, drawing on analytic theories of endorsement, relational normativity, and socially situated self-expression. On this basis, we argue that authenticity cannot be directly engineered by design, since it depends on reflective agency and context-dependent interpersonal obligations. Third, we advance a design-oriented proposal aimed at mitigating authenticity-related concerns without resorting to paternalistic restrictions: the introduction of optional and granular disclosure mechanisms ("AI ticks") that allow users to signal the presence and degree of AI assistance in a context-sensitive manner. We argue that such forms of transparency function not as a substitute for authenticity, but as a pro-ethical condition that enables agents to negotiate authenticity-based obligations within their relationships. The analysis contributes to current debates on the ethics of AI-mediated communication and the conditions of responsible human–AI interaction.

Keywords: Authenticity; Transparency; AI-Mediated Communication; Second-Person Perspective; Responsibility; Human–AI Interaction; Ethics of Communication.

1. Introduction. A New Communication Environment

The rapid development of artificial intelligence (AI) systems has begun to reshape interpersonal communication. Increasingly, people use generative systems to draft, refine, or translate messages, making human–AI co-authoring a pervasive component of everyday exchanges. Hancock et al. (2020) describe this phenomenon as AI-Mediated Communication (AI-MC): "mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals." While AI-MC spans a wide range of private and public contexts, this paper focuses on one-to-one personal communication, where relational expectations and norms play a particularly important role.

AI-MC offers clear advantages. Empirical evidence shows that it can increase efficiency, reduce interpersonal friction, and even foster the use of more positive emotional language, thereby improving perceptions of closeness and cooperation (Hohenstein et al., 2023). Unsurprisingly, a growing number of users report turning to systems such as ChatGPT for support in writing emails, text messages, and other everyday communications (Chatterji et al., 2025).

Yet these benefits coexist with significant ethical and psychological challenges. AI-assisted communication can undermine trust (Glikson & Asscher, 2023; Jakesch et al., 2019; Liu et al., 2022), obscure responsibility for harmful or deceptive messages, and enable new forms of cognitive outsourcing that may erode expressive capacities and critical thinking (Lee H. P. et al., 2025). It may also exacerbate communicative and epistemic inequalities, as not all users possess equal algorithmic literacy or the skills required to supervise AI outputs effectively.

Among these challenges, concerns about authenticity have recently gained increasing attention. Several authors argue that in certain interpersonal contexts, especially within intimate or emotionally meaningful relationships, communicating "through" AI may violate relational expectations or even obligations of sincerity and personal engagement (Battisti, 2025a; Lavazza, 2025; Robbins & Blundell, 2026). In these settings, what is at stake is not simply the semantic content of a message but the expression of the agent's will, effort, and relational commitments. Even in less intimate contexts, many forms of everyday communication rely on a "personal touch" that AI systems may dilute or replace.

These concerns are becoming more urgent as the technological landscape shifts. Until recently, to obtain AI advice on replying to an incoming message, users had to rely on external AI applications, manually copying content into a Large Language Model (LLM) and supplying contextual information. Today, AI is rapidly becoming an integral component of messaging systems themselves: embedded in keyboards, messaging apps, and soon in-device language models capable of accessing conversational histories and generating seamless replies. As AI becomes more convenient and frictionless, the temptation to employ it and, subsequently, begin to rely on it increases, and authenticity-related concerns become both more widespread and more difficult to detect.

This paper addresses a question that remains underexplored: Can authenticity concerns in AI-MC be mitigated through the design of AI-integrated messaging systems and, if so, how? We argue that, due to the moral and relational complexity of authenticity obligations, these concerns cannot be directly encoded into design constraints without imposing controversial normative assumptions on users. However, they may be mitigated indirectly through mechanisms that enhance transparency about AI involvement in communication.

More specifically, Section 2 outlines the ongoing technical integration of AI into messaging systems and its likely evolution. Section 3 examines authenticity as an ethical concern in interpersonal communication, showing why design alone cannot determine when authenticity is required. Section 4 develops the relationship between authenticity and transparency and introduces "AI ticks", a flexible transparency-by-design tool that allows users to disclose their use of AI in context-sensitive ways. We argue that such features, while not solving authenticity problems directly, can support users in negotiating authenticity-related expectations within their relationships. This approach offers a practical and actionable framework for developers to address a central ethical challenge posed by AI-MC.

2. Technical Implementation and Future Directions of AI-MC

The use of generative AI as a mediator of mobile communication follows Bar et al.'s (2016) model of technology adoption, from the initial introduction of AI agents to the public to their appropriation in support of personal communications (creolization). This was followed by technological repossession by suppliers, which has led to new implementations that capitalize on users' appropriations.

We structure the discussion around a compact taxonomy of AI-MC implementations as it has currently evolved (Table 1), in which we refer to the level of AI integration within apps as being application *agnostic* or application *specific*. We further distinguish AI-MC implementations according to their task specificity, i.e. the training of their agent model. As such, AI agents are considered as *general purpose* (not specifically created or trained to support communication) or *communication support* (created or trained specifically for that purpose). Finally, based on van Berkel et al.'s (2021) categorization of interactions with AI, we consider the interaction style with the AI agents as being *intermittent* (requiring a dialogical approach where the AI *offers* assistance only on user request) or *continuous* (the AI offers assistance continuously and without user instruction).

Table 1 A compact taxonomy of AI-MC implementations in the current technological landscape

	Level of Integration	
	Application Agnostic	**Application Specific**
Task Specificity	General Purpose, Communication Support	Communication Support
Interaction Style	Intermittent	Intermittent, Continuous

2.1. Application-Agnostic – General-Purpose AI for AI-MC

The availability of open-access generative AI tools has been rapidly appropriated by users to support communication-related tasks, ranging from translating emails to drafting highly tailored messages (Chatterji et al., 2025). The earliest and probably still dominant mode of AI-MC relies on *external prompting* of what we term here *Application-Agnostic* AI (AA): users switch to *General-Purpose* (GP) AI applications that are independent of the messaging environment (e.g. ChatGPT), manually supply context (incoming and outgoing messages as well as other relevant information and instructions), and then reinsert the AI-generated output into the conversation in the messaging app. According to van Berkel et al.'s (2021) categorization of human–AI interaction, this *intermittent interaction* paradigm follows a *dialogical* approach, where the human and AI system take turns in providing input (humans) and output (system).

AI assistance accessed through stand-alone apps or browser interfaces requires users to engage in a form of "on-the-fly prompt engineering". When a message arrives and the user needs help composing a contextually relevant and coherent reply, the assistant must be given explicit information about the sender, the relationship, the communicative goal, stylistic preferences, and contextual constraints such as tone or character limits. On mobile devices, writing such long, detailed prompts is far from trivial. Text entry is constrained by small screens, lack of tactile stability, and

distractions such as walking or typing on public transport. The cognitive effort involved in constructing a complete, well-specified prompt may be amplified when the user is already stressed, rushed, or emotionally affected by the incoming message. Appendix A (Supplementary Materials) illustrates the friction in a dialogical interaction with an AA-GP assistant.

2.2. Application-Agnostic – Communication Support AI for AI-MC

The public's appropriation of AA-GP agents to support communication has inspired some developers to repossess the technology (cf. Bar et al., 2016) through purpose-built, application-agnostic communication support agents (AA-CS). These constitute custom-prompted GP models, built specifically to support communication activities independently of the communication app. Services like "Rizz" (Rizz, n.d.) and "YourMove AI" are built on top of popular GP models to support online dating communications. They allow users to upload a screenshot of a chat from any dating app, and provide recommendations about how to reply to messages, based on the screenshot message content. All aforementioned tools require an application context switch, as do AA-GP implementations.

A further expression of AA-CS agents is AI tools built directly into smartphone keyboards. By integrating the AI agent within the keyboard, the AI is brought "inside" any application where it is required, avoiding the frustrating external or in-app application context switching. Examples of these integrations include virtual keyboards with AI capabilities, such as Samsung Galaxy Keyboard, Google Gboard, Microsoft SwiftKey and Apple iOS keyboard (see Figure 1).

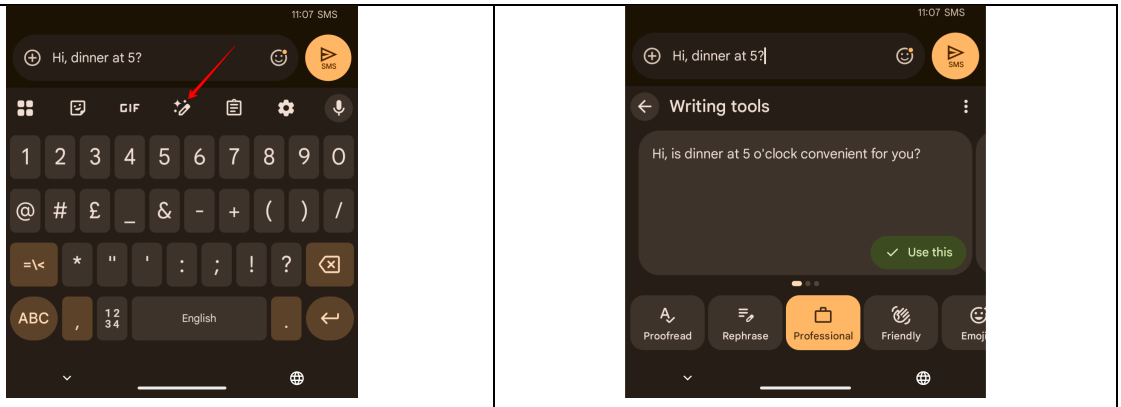


Fig. 1 The user invokes a text entry AI assistant from within the Gboard keyboard user interface, by pressing the "AI assistance" icon (left). The AI recommends linguistic modifications based on the original user input (right)

One major difference between AA-CS agents integrated with keyboards and AA-GP ones is that the former only offer pre-defined actions to the user, e.g. to proofread a draft or to change linguistic style by selecting from a list of predefined styles, such as "formal", "friendly", etc. (Figure 1, right). This saves the users from having to write detailed prompts on every occasion, facilitating access, but at the same time negating the user's ability to adapt the agent's behaviors with more detailed instructions or contextual information that might be relevant to the task. Another fundamental limitation is that the keyboard-integrated AI agent can retrieve *only* the text currently being composed in the messaging application, owing to the implementation of virtual keyboards as stand-alone apps, which do not have access to the data of other apps running on the device (see Appendix B for a detailed technical explanation).

2.3. Application-Specific – Communication Support AI for AI-MC

The final type of AI-MC agent is one that is integrated deeply into the application environment from which it is invoked (Application-Specific Communication Support, AS-CS). In this regard, we can further differentiate between AS-CS agents using van Berkel et al.'s (2021) typology of intermittent vs. continuous interaction. In the first subtype (intermittent interaction), humans provide instructions, inspect the AI-generated output, and provide further instructions as needed. An example of such AI-MC integration is Meta's embedding of an AI agent directly inside the WhatsApp platform (Figure 2, left), which allows users to modify draft text inputs for linguistic style (WhatsApp, n.d.). This approach works similarly to AA-CS keyboard integrations. However, because the agent is now integrated *within* the messaging application, the agent is not constrained by security requirements like keyboard integrations. It could, in principle, have full access to all conversation history, analyze it, extract contextual cues (e.g. linguistic patterns, conversation partner name, gender, relationship to user etc.) and use this information to dynamically augment the developer-programmed prompts behind every preset action, in order to adapt the output to the conversation and relationship context.

While it is not clear whether WhatsApp AI tools actually leverage conversation data, there exist other services that do. For example, the popular Viber messaging app has recently partnered with OpenAI (ChatGPT) to offer suggested replies to incoming messages (Figure 3), which can be generated without requirement for any manual text entry by the user. To accomplish this, it sends the latest 10 messages in the conversation to ChatGPT, as a means of providing context to inform the suggested reply generation (Viber, n.d.). Other platforms have also emerged as examples of AS-CS implementations. For example, services, such as Taplio (Taplio, n.d.) and Will (Will, n.d.), analyze a user's LinkedIn posts and auto-generate entire posts imitating the user's personal linguistic style and interests.

The second sub-type of AS-CS agents follows the paradigm of continuous interaction, where the AI system incessantly processes the user's input stream, and offers suggestions without the need for invocation by the user. The user can inspect, and ignore or accept these suggestions in real time (as they type). One of the earliest known AI-MC examples that follows this paradigm is Gmail's Smart Compose in-line smart sentence completions (Figure 2, right) as described in Chen et al. (2019). A predictive text engine attempts to offer plausible sentence completions to users as they type, although the user has no control over the behavior of this mechanism (e.g. to select linguistic styles, or to tune other operational parameters). Currently, these tools are not highly personalized, in the sense that they do not explicitly pick up on users' individual personal data histories beyond the current email thread, and therefore cannot provide fully contextualized outputs.

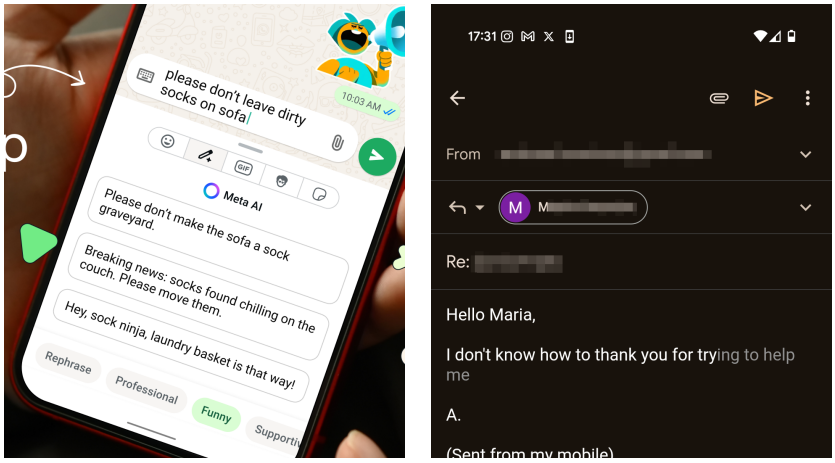


Fig. 2 Sample integrations of AI assistance within the confines of individual messaging applications. On the left, the AI offers linguistic restyling options based on the original user's input (WhatsApp Writing Tools). On the right, the AI offers sentence completions (grey characters) that follow the user's original input (Smart Compose for Android)

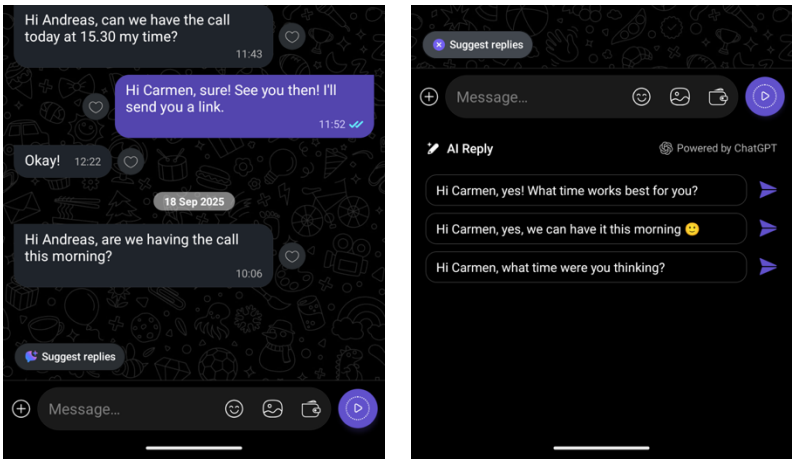


Fig. 3 Sample integrations of AI assistance in the Viber app

Across these approaches, we observe a clear trajectory: assistance becomes more immediate, prompting becomes more invisible, and interaction requires less deliberate user effort. Trust in the system behavior is afforded by allowing users ultimate control over the acceptance of generated text, following Shneiderman's (2021) *Human Control* mantra in AI systems interactions (*Preview first, select and initiate, then view execution*); such designs have the potential to greatly facilitate human–AI collaborative writing tasks.

Summarizing, we discuss how core attributes characterize the various types of AI integration (AA-GP, AA-CS, AS-CS, and the prospective CAA-CS architecture introduced in Section 2.4). These attributes and the level at which they are exhibited are mapped in Table 2, expanding on our compact taxonomy of AI-MC tools. A detailed description of these attributes can be found in Appendix C (Supplementary Materials).

Table 2 Core attributes in the expanded AI-MC tool taxonomy and their level of presence (None, Low, Moderate, High, Full). Integration is (E)xternal, (I)n-app, (K)eyboard. Functions are (R)estyling, (G)eneration, (C)orrection, Co(M)pletion

AI-MC Architecture	**Integration**	**Function**	**Contextual Access**	**User Control**	**Prompting Effort**	**Personalization**	**AI Contribution**	**Interaction Paradigm**	**Examples**	**Availability**
AA-GP	E	R, C, G	None	High	High	None	Low to Full	Intermittent	ChatGPT, Meta AI	Current
AA-CS	E	G	None	High	None	None	Full	Intermittent	Rizz, YourMove AI, Lavender.ai	Current
	K	R, C	Low	Moderate	Low	None	Low to Moderate	Intermittent	Google Gboard, Apple iOS keyboard	Current
AS-CS	I	M	Low to Moderate	None	None	Moderate	None to Moderate	Continuous	Smart Compose, Hyper Write	Current

	I	R, C	Low	Moderate	Low	None	Low to Moderate	Intermittent	WhatsApp	Current
	I	G	High to Full	None	None	None to High	Full	Intermittent	Viber, Gmelius, Taplio, Will	Current
CAA-CS	K	R, C, G, M	High to Full	Low to High	None to Low	High	Moderate to High	Intermittent, Continuous	N/A	Proximal Future

2.4. Moving Beyond Assistance: Contextual and Personalized Co-authoring

Due to the proprietary application code, it is not known how individual conversation context is considered in extant implementations of the AS-CS approach (i.e. what prompting instructions are sent to the AI model along with conversation data). Yet, if the tighter integration of AS-CS agents with application data were to be exploited, AI agents could easily shift from being largely responsive (refining or continuing what the user has already typed) to becoming more proactive. Messaging applications already contain rich streams of contextual information that could, in principle, be used to automate prompt construction and reduce user workload even further, as demonstrated in the case of Viber. While Viber's capabilities are application-specific (i.e. available only within the app), it is perfectly technically feasible for application developers to expose conversation-related message histories through programming interfaces (APIs) to other apps, including the keyboard. It is also technically feasible for operating system developers (e.g. Android, iOS) to modify existing keyboard-input field bindings so that keyboards can expose details of how text is generated (i.e. by the user, by an AI, or both), so that the application invoking the keyboard can receive this information (a possibility directly relevant to the platform-recorded disclosure discussed in Section 4.1). As such, it is possible to imagine a very proximal future, where this two-way data exchange between keyboard-integrated AI agents and messaging applications could give rise to a Contextual Application-Agnostic Communication Support architecture (CAA-CS), enabling a more advanced stage of AI-MC that involves *contextualized prompting*, where the assistant automatically infers relevant cues from the ongoing conversation, mining information autonomously off-line or in real-time.

If such cues are made available to the assistant, the user may no longer need to specify detailed instructions. Let's return to the example of Miranda (Appendix A), and imagine she is having a slightly different conversation with her boss, about the delivery of a report (Figure 4). In response to his question about whether the report can be done by Monday, she could simply write:

"can't b4 wedsday @ai."

In this example (Figure 4, left), use of a "tag" (@ai) could invoke the contextualized AI assistant, which could automatically parse the entire conversation history, recognizing from her past interactions that their relationship is professional rather than informal. It would then adopt an appropriately formal tone, omit emojis that she rarely uses in work-related exchanges, generate a reply length consistent with her typical SMS format, correct any spelling errors ("wedsday") and include or omit an apology depending on whether such expressions usually accompany her communications in similar circumstances. A contextually grounded, highly personalized AI-generated reply, such as *"I'm sorry, Chris, John will have to wait for that report until Wednesday."*, could therefore emerge from a minimal-effort user instruction, by dynamically augmenting a base prompt with the user's input and contextual information at runtime (Figure 4, center). Miranda could be afforded high control over the generation process, by instructing re-tries, or manually picking the language tone. Contextual real-time text suggestions could also be offered (Figure 4, right) with similar levels of control over the generative process as the intermittent interaction example.

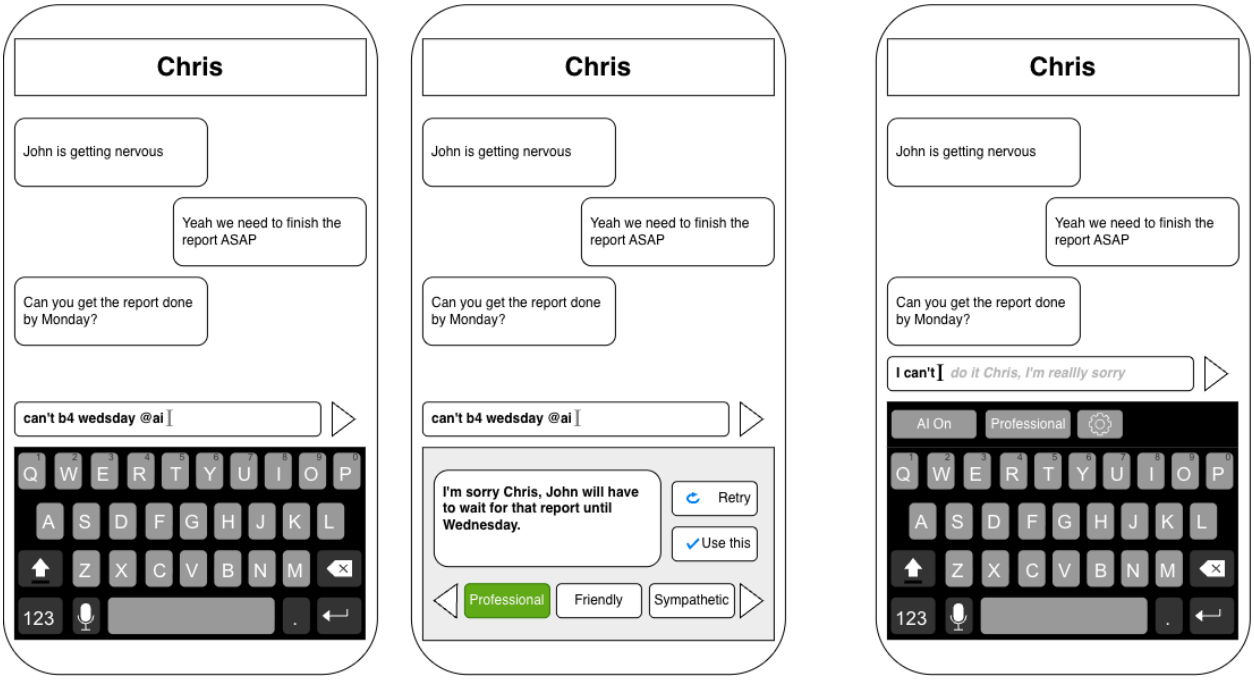


Fig. 4 A fictional conversation between Miranda and her boss, assisted by a CAA-CS AI agent integrated in the keyboard. Miranda invokes the agent by tagging it while writing a short input message (left). The AI agent automatically infers the appropriate language tone and generates a likely reply, integrating information from Miranda's brief input and the conversational context (center). Real-time suggestions as the user types can also be offered, with control settings over enabling assistance and assistance operational mode (tone) easily available on the keyboard UI (right)

However, as contextualized prompting becomes feasible, the nature of AI-MC changes qualitatively. The assistant no longer merely edits or expands explicit user input; it becomes a *co-author* capable of reconstructing communicative intentions from context and history.

With deeper integration come new technical, legal, and ethical questions. If AI assistants are to infer conversational context automatically, they may need access to large amounts of conversational data, either during training or at the time of use. This raises immediate issues: How can such data be collected ethically at scale? How can systems avoid embedding cultural, linguistic, or relational biases? Are such biases intrinsically undesirable in mediated communication, or are they even appropriate under specific contexts? How can users retain control over what their devices analyze, store, or transmit?

Users may also want to understand or correct the contextual cues that systems infer. If a messaging platform mistakenly interprets an acquaintance as a close friend, or misidentifies the appropriate tone for a conversation, should users be able to inspect and modify these inferences? Should contextual cues added to the linguistic modification or generation prompt be reviewable, even if only in an optional transparency panel?

More deeply, contextual generative systems raise the question of authenticity. If the AI produces a message that mirrors the user's typical style and expresses the sentiment they would have articulated, is the resulting communication authentically theirs? Or is it fundamentally authored by the system? And if the recipient cannot distinguish between the two, to what extent does AI reshape relational expectations?

Finally, the prospect of interoperable AI agents that negotiate tone, intent, or emotional valence across interlocutors (i.e. both parties using AI) introduces further normative concerns: Would such agents compromise autonomy by smoothing disagreements or pre-emptively removing potentially sensitive phrasing? How would such mediation affect responsibility, trust, or the perceived sincerity of interpersonal exchanges?

At the current state of the art, AI does not yet demonstrate intentionality or a will to act upon any self-formulated goals (Ngaihlian, 2025; Redaelli, 2025), although as systems become more autonomous, it might be increasingly pertinent to decouple AI agents' morality (and ability to answer these questions) from that of its designers or users. In the context of AI-MC therefore, it is probably premature to speak about an ethical AI operating on behalf of the users, but it is appropriate to discuss answers to the aforementioned questions as a matter to be resolved through ethical design, or ethical use of the AI agent. Thus, while these trajectories do not determine how future AI-MC will be deployed, they show that the design decisions made today will shape the communicative landscape of billions of users. Default settings – typically accepted by most users – will exert a disproportionate influence on emerging norms. And given that developers cannot be expected to anticipate all implications of relationally embedded AI, questions of authenticity become central in evaluating the ethical significance of these technologies. We now turn to authenticity as a key moral dimension of AI-mediated communication.

3. Authenticity and AI-Mediated Communication

AI-mediated human communication has already sparked a debate about the authenticity of communication, particularly concerning the integration of AI into messaging systems.

Across experimental, qualitative, and design-focused work, empirical evidence shows that authenticity emerges as a central tension in AI-mediated communication. A preregistered experiment shows that explicitly disclosing AI involvement substantially lowers perceived authenticity and sincerity, whereas leaving AI use uncertain preserves impressions comparable to (or sometimes better than) clearly human-written messages (Zhu & Molnar, 2025). Diary and interview studies of everyday interpersonal use find that while assistance boosts confidence and helps people find precise language, it often yields verbose, unnatural, or overly emotional phrasing that feels inauthentic and raises worries about overreliance, with acceptability varying by stakes and relationship formality (Fu et al., 2024). Interviews with Black users document direct authenticity harms when systems "correct" African-American vernacular English or names – producing exclusion, mistrust, and concerns about cultural erosion – and argue for consentful, culturally responsive personalization to preserve voice (Basoah et al., 2025). A design-space review codifies authenticity/ownership as a core user–system dimension, noting that perceived authenticity tends to diminish as the system contributes more text and recommending steerable personalization and transparency to mitigate this trade-off (Lee M. et al., 2024). These findings are further supported by more recent evidence showing that heavy AI assistance in drafting intimate messages consistently reduces perceived ownership and authenticity of the message, with gains in clarity and competence failing to offset this loss of authenticity (Fan et al., 2026).

Although empirical evidence highlights the existence of authenticity-related tensions, the underlying ethical sources of these tensions remain relatively under-explored. It is not yet clear, in fact, which understanding of authenticity is morally relevant to this debate. For this reason, it is necessary to provide a conceptual framework of authenticity within the recent literature and then develop its normative implications.

3.1. Authenticity as a Multidimensional Concept

According to Lipovetsky (2021), authenticity denotes the contemporary injunction to "be oneself," understood as the free and sincere expression of individuality without external masks or constraints. In its current form, authenticity has been sacralized as a collective value while also becoming a cultural fetish, marketed through lifestyles, consumption practices, and experiences that promise to reveal the "true self." At the same time, it has assumed a normative dimension, functioning not only as a personal aspiration but also as a relational obligation to present oneself as authentic – and, conversely, to expect others to present their authentic selves.

While the contemporary debate on authenticity spans multiple philosophical traditions, including continental phenomenology and existentialism (Varga & Guignon, 2023), this paper focuses on several influential analytic and dialogical accounts that help clarify why authenticity in communication cannot be reduced to stylistic similarity or behavioral predictability.

A first family of views conceives authenticity in terms of reflective ownership (or endorsement). On such accounts, associated primarily with Frankfurt (1971, 1988) and the reflective turn in moral philosophy, an expression is authentic

when it reflects attitudes or commitments that the agent identifies with upon reflection (e.g., Korsgaard, 2009). From this perspective, what matters is not whether a message resembles the agent's usual style, but whether the agent can recognize it as something they would endorse.

A second family of views highlights the relational dimension of self-understanding and agency. Although Velleman (2005) does not explicitly develop a theory of authenticity, his narrative account of the self provides an important structural basis for understanding how sincerity and self-expression are shaped by interpersonal norms. On Velleman's view, agency emerges through practices of self-ascription and commitment that are always situated within relations to others. What counts as an appropriate or sincere self-description therefore depends in part on the expectations that hold within a given relationship: greater openness is appropriate in intimate contexts, while restraint is required in more formal ones. From this perspective, an utterance may be linguistically correct and yet inauthentic if it fails to respect the normative expectations of the relational context in which it is produced.

Finally, authenticity has been understood as a socially situated practice. Taylor (1992) argues that authenticity is not merely inwardness but a dialogical achievement, negotiated within shared forms of meaning that constitute a moral horizon. On this view, authenticity depends on how agents locate themselves and their self-interpretations within social practices, roles, and the social goods they define. AI-mediated communication can reshape these practices by redistributing communicative labor or obscuring the user's contribution to an expressive act, thereby altering the conditions under which authenticity is constructed.

Taken together, these perspectives show that authenticity is multidimensional, involving reflective endorsement, relational norms, and social context. It should nevertheless be distinguished from several related concepts. To be sincere is to express attitudes or feelings one genuinely holds, while honesty more broadly requires avoiding deception. Authorship identifies who formulated or produced a message; ownership, as we use the term here, indicates whether senders recognize and endorse the message as their own. Agency captures their intentional participation and control in producing and sending it, whereas responsibility refers to their answerability for the communicative act and its foreseeable effects.

These dimensions do not always coincide. A message drafted by AI may sincerely express an attitude endorsed by the sender, who may remain responsible for sending it even without being the author of its exact wording. Conversely, a wholly human-authored message may be insincere or inauthentic. Authenticity, by contrast, connects an expression with the agent's practical identity and with the legitimate expectations arising within a relationship. This account provides the basis for the notion of second-person authenticity developed specifically in the debate on AI-mediated communication.

3.2. Second-Person Authenticity and AI-MC

In order to introduce the concept of second-person authenticity and illustrate how it can be violated through the mediating role of AI, it is helpful to consider communication within a romantic relationship, especially when it comes to delivering particularly important messages, such as apologies. Battisti (2025a) has argued that massively relying

on ChatGPT to apologize to your partner can be morally problematic. This is because, in a relationship, there are non-transferable reciprocal duties that require a partner to personally perform certain tasks, which are called *authenticity-based obligations* (Battisti, 2025a). For instance, when you come home sad because your boss mistreated you, you generally want *your* partner – not someone else – to comfort you, regardless of whether others may be more skilled at it, for example a therapist. Likewise, after a quarrel, you want *your* partner to apologize.

Authenticity-based obligations rest upon the idea that the agent performing the task cannot be replaced by an alternative agent without the outcome of the performance being denuded of its moral significance. They differ from other kinds of obligations within relationships that do not require non-substitutability of the agent, namely, *outcome-based obligations.* For instance, when in a relationship you take responsibility for paying the electricity bill, your obligation to your partner is generally understood not as requiring that *you* perform the very act of payment, but rather that the bill be paid. This act can be delegated to a secretary or even an AI agent with no moral implications. What matters is the result rather than the agent who actually performs the action.

While outcome-based obligations undoubtedly serve pragmatic purposes, authenticity-based obligations hold particular significance: they enable partners to recognize and acknowledge each other as the distinctive persons with whom they deliberately chose to engage; indeed, the fulfillment of such obligations is a paradigmatic moment in which the bond of trust and intimacy is validated by the performative act of the other party. Authenticity-based obligations represent a fundamental component of the reciprocal obligations within a relationship precisely because they allow for a reflective consolidation of the relationship in itself, a necessary step for its maintenance and continuation over time, as long as the partners decide to continue to stay in that relationship.

The fulfillment of authenticity-based obligations is evaluated through the lens of second-person authenticity. According to Battisti (2025a), this form of authenticity is recognized by the other party in the relationship (e.g., one's partner) as long as: (a) the agent is the author of a given action, which implies directly performing it, namely undertaking the cognitive and emotional process required to bring it about; (b) the resulting action or behavior is consistent with the agent's practical identity as it has developed within the relationship with the partner. In other words, using an LLM to apologize on your behalf prevents you from actually performing the act of apologizing and from engaging in the cognitive and emotional labor necessary for the other party to validate the bond of intimacy and trust, through which your partner recognizes you as the person they are in a relationship with.

While in some contexts the violation of second-person authenticity in AI-MC may be quite clear, in many cases it is more nuanced. First, not all uses of LLMs undermine second-person authenticity: for instance, when a couple does not share the same language, using AI solely to translate text may not only avoid undermining any authenticity obligations but may even help fulfill them more effectively in some circumstances. Moreover, what happens when, instead of massively relying on an LLM to communicate, we simply use it for editing, integration, or adjusting the tone? As Lavazza (2025) points out, there may be gray-zone circumstances in which AI-mediated communication to fulfill authenticity-based obligations is not massively outsourced to the language model but rather *co-authored*, which still allows the partner to engage in the cognitive and emotional process. In these cases, it is difficult to claim that there is no at least partial authenticity commitment on the part of the sender, even though – from a practical point of

view – it would remain difficult for the partner to qualitatively recognize the extent of the agent's contribution, thus hindering the possibility of the reflective consolidation of the relationship discussed above (Battisti, 2025b). Others, such as Fröding and Peterson (2025), have cast doubt on the general thesis on authenticity, stressing that authentic communication does not require the performative effort of apologizing but rather the genuineness of one's intentions.

It should further be noted that individual relationships develop their own distinctive features. Consequently, what might typically be understood as an outcome-based obligation could, depending on the relational context and the partners' shared understanding, be reinterpreted as an authenticity-based obligation. This calls for a contextualized, case-specific assessment. While above we discussed examples such as apologizing to or consoling a partner – situations in which it is plausible that many people would expect the partner personally to perform the act in question, so giving rise to authenticity-based obligations – other couples might attach authenticity-related value to different tasks, such as doing the grocery shopping together every week.

In this paper, we limit ourselves to highlighting the existence and moral importance of authenticity-based obligations, providing a few paradigmatic examples, without committing to the claim that some tasks within a relationship necessarily require authenticity while others certainly do not. In this way, we remain open to the possibility that there may be other contexts beyond the example we provided in which a particular couple attributes authenticity value to specific tasks that other couples generally do not.

A further layer of complexity should be highlighted. Although this phenomenon has been discussed mainly in romantic relationships, it remains an underexplored question whether authenticity concerns may arise in other types of relationships as well. Some possible candidates are relationships between relatives and close friends, which in some respects may resemble romantic ones, as well as professional relationships, such as those between colleagues or between doctors and patients.

To determine whether authenticity-based obligations exist in such contexts, one must ascertain whether they legitimately apply to these types of relationships as well. Addressing this issue would require an in-depth analysis of each type of relationship and therefore falls outside the scope of the present work.

However, we should recognize that a bond of trust and intimacy – the element that justifies authenticity-based obligations in a relationship – also plays a significant role in non-romantic relationships. This is most evident in friendships and family ties, yet it also emerges in other spheres, such as the workplace. This latter case deserves particular attention. Professional relationships are clearly different from romantic relationships, friendships, or family relationships, as they are more evidently governed by external norms. They also often lack the initial choice or morally significant voluntary commitment involved in entering a specific relationship, given that colleagues are frequently assigned rather than chosen. Even so, conditions can still arise for an additional layer of connection based on mutual trust and a greater or lesser degree of intimacy. This goes beyond mere adherence to minimum workplace standards of propriety, which do not require assessment through the lens of authenticity because other mechanisms regulate them; indeed, no trust or intimacy is assessed in that regard. Nevertheless, trust and intimacy can give rise to legitimate authenticity-based obligations, which serve precisely to test that further layer of the relationship.

Consequently, if authenticity-based obligations can legitimately emerge in other contexts, a fundamental theoretical task is to identify the conversational characteristics that signal their presence or absence across relationship types and environments, without limiting the analysis solely to romantic connections. To this end, we propose two criteria, which are not meant as necessary or sufficient conditions, but as useful indicators for assessing obligations based on authenticity.

A first potentially valuable criterion may concern the extent to which communication involves intimate content conveyed by one party to the other by virtue of the trust that bonds these relationships. For instance, if a colleague decides to talk to me about something very personal and reserved – for example, the possibility of changing jobs or mobbing issues – it is likely that they want a response from me, considering the trust they place in me. It is doubtful that they would accept a message filtered in form and content by an AI, which would prevent the bond of trust between us from being assessed, unlike through a performative act of responding, which involves cognitive and emotional labor.

Another criterion may concern whether communication is aimed at transmitting specific operational information. For example, when a doctor asks to have an email revised by AI to make medical language more accessible when communicating a diagnosis, the use of AI might even enhance the clarity and quality of the relationship with their patient. When communication is limited to the conveyance of purely practical content, one would not even expect one's romantic partner to avoid using AI, for instance when they want to explain condominium waste regulations to me.

This non-exhaustive discussion allows us to understand that authenticity can be a requirement in some communications and not in others, across different relationships, something that further increases the moral complexity of authenticity-related constraints in relational contexts. Moreover, it allows us to identify who can legitimately hold authenticity expectations, namely those who share a relationship of intimacy and trust with another person, and under what circumstances, namely when intimate communications are at stake rather than merely operational ones. However, this deliberately remains a general framework. As mentioned, obligations grounded in authenticity depend on factors distinctive not only to the type of relationship, but also to each specific relationship. When it comes to these obligations, each relationship possesses its own moral framework, which may overlap with those of other relationships in some respects but may also differ in substantive ways.

Finally, it should be considered that there may be cases in which one partner believes that a task ought to be performed as an authenticity-based obligation, while the other disagrees. This may give rise to unreasonable requests, even within the type of relationship that would otherwise permit such requests. Precisely because of the deeply personal nature of these obligations, it is difficult to establish when expectations are or are not reasonable. In this section, we have offered some elements that can guide partners in defining their own moral framework, but we acknowledge that further research is needed to distinguish legitimate from illegitimate expectations.

As a general guideline, legitimate authenticity expectations should be grounded in the nature and history of the relationship, be proportionate to the significance of the communicative act, and avoid imposing undue control over the other person. For instance, expecting a personal apology may be reasonable, whereas demanding disclosure of every minor form of linguistic assistance may not be.

3.3. Can Authenticity Be "Enforced" by Design?

In light of this reconstruction, one may ask whether it is possible, or even desirable, to address authenticity issues directly through the design of AI-integrated messaging applications. We remain skeptical about this possibility. Because of the complexity of relational contexts, we believe it is doubtful that designers could anticipate and prevent the emergence of such problems through design alone. Precisely because these are moral issues embedded within specific relationships, it would be virtually impossible to predict in advance the contexts in which they may arise.

Moreover, even assuming that some of these contexts could be identified – for instance, AI could recognize that a specific conversation involves potentially sensitive topics based on the criteria outlined above – imposing a particular value framework on users would remain ethically problematic, especially in the pluralistic environments in which such applications are deployed. In sum, embedding authenticity-related concerns directly into design is highly problematic or, at best, unfeasible.

This conclusion does not in any way mean dismissing authenticity-related problems, quite the opposite. A closer look reveals that such problems become even more pressing when moving from external software not integrated with the messaging system to an AI integration directly embedded in messaging platforms, as described in Section 2. The former case includes, for instance, copying and pasting sentences from an LLM, as already discussed in previous literature. The growing accessibility and ease of use will make authenticity-based concerns even more salient and pervasive, also due to the structural human tendency to convenience.

Nevertheless, avoiding the development of such systems would prevent people from enjoying the benefits of AI-mediated communication we mentioned. Such a restriction, wherever feasible, would be overly demanding and paternalistic. Instead, what is desirable is that users become aware of authenticity issues and engage in discussion with their romantic or professional partners to define their mutual expectations and obligations regarding AI-mediated communication.

Therefore, although authenticity cannot be specified at the level of messages or implemented directly through system design, at the same time communication technologies must avoid undermining the conditions for authenticity and support the user's ability to occupy an authentic stance within relationships. Although we have discarded the possibility of *directly* intervening to prevent authenticity-related problems, we believe that there may be indirect and non-paternalistic ways to address the issue of authenticity by design. This concerns transparency and disclosure: while these are conceptually and ethically distinct from authenticity, they can nonetheless play an important practical role in addressing authenticity-related concerns in AI-mediated communication. To this end, in the next section we will

present the issue of transparency, its relationship with authenticity, and then suggest a practical recommendation regarding the disclosure of AI-MC use, which may indirectly prevent or mitigate authenticity-related problems.

4. Transparency as a Practical Tool

Transparency refers to the possibility of accessing information, intentions, or behaviors that have been intentionally revealed through a process of disclosure (Turilli & Floridi, 2009). Being transparent about relevant information in a conversation allows the interlocutor to make certain decisions – or behave in certain ways – that they otherwise would not.

For instance, if an influencer promoting a commercial product through their channels does not disclose that they have been paid to do so, their followers may fail to properly evaluate the product's quality. Similarly, if John hides from his vegetarian partner the fact that he works in an industrial livestock farm, this could have consequences regarding whether they decide to stay with him or not.

Transparency is often not regarded as an ethical principle *per se*. In the field of information ethics, Turilli and Floridi (2009) consider transparency as a *pro-ethical condition*, that is, a condition that enables the promotion of higher-order ethical principles. In a similar vein, in the field of journalism ethics, Craft and Vos (2021) think of transparency as a means to serve ethical imperatives such as accountability and truth telling. Accordingly, in the previous examples, adopting transparent behavior serves to promote honesty and the interlocutor's decision-making autonomy.

The analysis developed in Section 3 suggests that authenticity in communication is not a property of messages themselves, nor a psychological state that can simply be detected or reconstructed by technical systems. Rather, authenticity depends on the user's endorsement of what is expressed, on relational expectations between interlocutors, and on socially situated practices of self-presentation. These features are normative, context-dependent, and irreducible to design parameters.

Given this structure, AI-MC systems cannot meaningfully guarantee or enforce authenticity. What they *can* do, however, is avoid undermining it. This shifts the design focus from shaping the content of communication to shaping the informational and relational conditions under which users can authentically position themselves, most notably through forms of transparency and disclosure.

In light of this understanding, for the purpose of this paper, it becomes interesting to explore the relationship between transparency and authenticity to see what happens when, in communications that involve authenticity requirements, the partner using AI-MC decides to inform the other partner about this use.

One might argue that, in doing so, authenticity problems could disappear since the partner demonstrates honesty in disclosing the use of AI. In light of the evidence mentioned above, people tend to recognize a solid connection between transparency and authenticity. Undisclosed AI assistance may indeed undermine trust and credibility, especially if recipients assume that messages reflect unmediated personal thought. Although this claim may grasp something relevant to the discussion, there are reasons to be skeptical about this. This is because authenticity, in the sense developed in the previous section, cannot be simply reduced to the notion of honesty.

A person may be perfectly honest about the means used to compose a message, and yet that message may still turn out to be potentially inauthentic. To better explain this point, let us return to the example of a partner who apologizes to the other; if the partner uses an LLM and discloses this to the other, it seems consistent with our moral sense that the latter might still be dissatisfied. Indeed, authenticity concerns do not stem from the mere lack of disclosure *per se*, but from the still present impossibility of assessing whether the person performing the apology actually meets the requirements suggested by second-person authenticity, which is functional to the reflective consolidation of the relationship.

The inauthentic apology case is more similar to the issue of transparency in the case of a student who uses cheat sheets during an exam (Battisti, 2025a). Here, what is being concealed is *in itself* morally wrong, regardless of whether it is hidden or revealed. The wrongness does not lie solely in the act of hiding the cheat sheets, but primarily in the unfair advantage the student gains over their peers and in the teacher's inability to accurately assess the student's actual level of preparedness. Similarly, in the case of a partner apologizing through an LLM, what is problematic is not so much – or at least not only – the fact of concealing from the partner the use of the LLM, but the very use of the tool itself. In this sense, transparency regarding the use of LLMs is not a sufficient condition for directly avoiding authenticity concerns in interpersonal communication.

From a more practical perspective, transparently disclosing the possible use of an LLM in crafting a message may still assist the partner in understanding whether an authenticity requirement has been violated or not. Disclosing the use of AI may, in fact, be crucial in several ways to deal with authenticity. The most important is that if a person is required to communicate when they make use of AI integration, then by not doing so – by communicating without AI assistance – they may signal their intention to respect authenticity-based obligations, although such a signal cannot guarantee that these obligations have been fulfilled.

Furthermore, disclosing the use of an LLM may also be relevant in more complex cases than those discussed above, namely, when a partner uses an LLM not by relying on it extensively, but as a supportive tool to find the right words to meet authenticity requirements, either because they feel insecure or because they want to learn to better fulfill those requirements (Lavazza, 2025). Although there might be reservations about claiming that a person using an LLM in this way fully satisfies the authenticity condition, the very act of disclosing the AI use and the reasons behind it to the partner may nonetheless be valuable for the couple, demonstrating greater commitment and willingness to initiate a process that could eventually lead to more authentic behavior in the future. Through disclosure, the partner may assess second-person authenticity not immediately but later, taking into account the commitment the other partner shows and their willingness to learn how to apologize more genuinely (Battisti, 2025b).

Therefore, transparency can be understood as a requirement that, although not sufficient to resolve authenticity concerns, may nonetheless indirectly mitigate them simply by signaling to one's interlocutor that AI has been employed in situations where there is a legitimate expectation that it should not be.

4.1. How to Implement Transparency by Design

If we recognize transparency as the pro-ethical condition that indirectly helps mitigate authenticity concerns in the context of integrated AI in messaging apps, we must now investigate, practically speaking, to what extent such transparency should be promoted by design. In what follows we discuss two strategies.

The first strategy suggests that designers include in messaging applications a feature that always notifies the recipient when the interlocutor is using AI. As soon as the user activates the AI integrated into the keyboard, a short notice such as "generated with AI assistance" could automatically appear within the message for the receiver. In other words, disclosure is based on AI assistance recorded by the platform and is compulsory at the system level and not controllable by users. Thanks to this strategy, it becomes possible for the interlocutor to know whether AI is being used in conversation, and this may allow for greater control within the relationship regarding the fulfillment of authenticity-based obligations. Moreover, such a system would encourage the speaker to avoid using AI when they are aware that the other party has expectations that AI should not be used, thereby indirectly preserving the second-person authenticity requirement.

However, this recommendation can be criticized as too demanding, as it would impose an unfounded obligation to disclose AI usage in *every* conversation. In fact, as mentioned earlier, in many contexts the use of AI may not be ethically problematic from an authenticity point of view, while at the same time the interlocutor may prefer not to reveal that AI assistance has been used. This preference may be justified by the fact that they are not undermining any authenticity-based obligations. This may be either because they are in a type of relationship in which such obligations do not arise, as in a standard relationship between a buyer and a seller where operational and non-intimate information is exchanged, or because not all uses of AI are problematic, not even within relationships that might require second-person authenticity in specific circumstances. With regard to the latter situations, consider this example: in general few would consider it ethically necessary to disclose the use of human editors or translators when composing a letter, as conditions (a) and (b) for second-person authenticity are met; however, if the receiver were to learn that a letter had been drafted with AI assistance in such a context, when in fact it had merely been translated or grammar-checked, they might mistakenly believe that it had been written entirely, or to a large extent, by the AI.

Moreover, from a practical perspective, making disclosure mandatory could discourage many individuals from using messaging systems that incorporate such technologies. Rigid disclosure norms may seem artificial or burdensome, especially given how deeply algorithmic aids are embedded into mobile communication tools. Therefore, although this strategy is in principle the best option for securing authenticity through transparency, it comes with moral and practical costs that are simply too high. This leads us not to endorse it.

While previous work has suggested that disclosure may contain detailed information, such as the history of edits or of AI use, which can offer transparency during communication (Komninos, 2024), such detailed information may not be practical in the context of fast communication between parties. Therefore, an alternative solution is more promising, one that may resemble and further expand the well-known "blue ticks" feature in messaging services such as WhatsApp or Telegram. These signal that a message has been delivered and read by the recipient. Let us call their AI-

MC equivalent "AI ticks." In this scenario, users deliberately communicate or omit information about AI use, depending on the context and the expectations shared among partners.

AI ticks could function unilaterally, insofar as the sender decides whether to enable AI disclosure or not. Alternatively, they may operate through a reciprocity mechanism. Partner A may choose to disclose to partner B the potential use of AI in their messages; if partner B also chooses to disclose their own potential use of AI, reciprocal disclosure becomes active. Otherwise, neither A's nor B's potential use of AI will be visible to the other. In other words, AI ticks work analogously to a tit-for-tat mechanism, which avoids mandatory disclosure as in the first strategy while allowing for a more balanced distribution of the costs of transparency. A further alternative is that transparency is governed through a negotiated mechanism. Users can set preferences for any contact or category of contacts, such as "disclose that I use AI," "I want to know if the sender uses AI," "I do not want to know," or "at the sender's discretion."

Importantly, AI ticks should not be understood as a mechanism that guarantees truthful disclosure, but as a tool that facilitates voluntary transparency. Users may still conceal their use of AI (for instance, by relying on external AI systems whose use the platform cannot record), and the absence of an AI tick should never be taken as proof that no AI assistance was involved. This limitation is an unavoidable consequence of preserving users' autonomy rather than imposing mandatory disclosure.

Table 3 Illustrative typologies of "AI ticks" for disclosure in AI-mediated communication[1]

AI-Tick Modality	Core Mechanism	Who Controls Disclosure	Typical Use Context
Unilateral disclosure	The sender independently decides whether to signal the use of AI assistance in composing a message.	The sender alone.	Relationships in which one party wishes to adopt transparency without prior coordination.
Reciprocal disclosure (tit-for-tat)	AI disclosure becomes visible only if both interlocutors agree to disclose their own potential use of AI assistance.	Both participants jointly.	Relationships structured around symmetry and shared expectations of mutual transparency.
Negotiated / preference-based disclosure	Disclosure rules are defined in advance through user preferences for specific contacts or relational categories.	Each user through configurable settings.	Contexts involving stable relationships with established norms and differentiated expectations.

[1] In all modalities, the tick may report AI assistance recorded by the platform or voluntarily declared by the user; see Section 4.1'

These three modes of AI ticks summarized in Table 3 may have different ethical implications regarding, for example, reciprocal expectations, the pressure to conform with others' requests – especially when it is unclear whether a given request is legitimate – and the potential strategic use of this mechanism.

Further ethical implications arise from the default setting that messaging applications adopt to manage AI disclosure. For instance, some applications might generate AI ticks by default whenever a user employs the platform's own AI assistance, requiring users to change their settings if they do not wish to disclose such use; others, by contrast, might leave AI use undisclosed by default. Default settings have relevant implications as they may affect users' behavior and shape reciprocal expectations. For example, default AI disclosure could lead individuals to systematically disclose their use of AI, thereby inhibiting legitimate uses of AI and potentially exposing users to undue pressure.

Addressing this issue goes beyond the scope of this paper and would require experimental studies to assess which default setting is most appropriate. We limit ourselves to highlighting this tension, while noting that, since AI disclosure for the fulfillment of authenticity-based obligations concerns a relatively narrow category of communications and a limited number of users, a less costly and more promising hypothesis to test may be to leave AI use undisclosed by default. Users could nevertheless manually adjust their settings when they wish to declare that they are not using AI, to request disclosure for a particular part of a conversation or for a specific contact, or to negotiate mutual disclosure.

Beyond reciprocity mechanisms, a binary tick showing only whether AI was used would not be enough. AI assistance comes in different degrees. Accepting two words suggested by autocomplete is not the same as using a complete AI-generated response. Translation, stylistic revision, and extensive human editing fall somewhere in between. AI ticks should therefore use practical categories that users can easily understand, rather than attempt to measure authorship precisely. A workable system might distinguish between linguistic assistance, such as spelling correction, grammar checking, and translation; the reformulation or completion of text written by the user; and the generation of substantially new content. A simple indicator could provide this basic information, with further details available if users wish to see them. This would allow for greater precision without making communication unnecessarily burdensome.

What can be disclosed also depends on where the AI was used. A messaging platform can record the use of its own AI tools, the kind of change requested, and whether the user accepted or later edited the suggested text. However, it cannot reliably identify text generated by an external AI system and then pasted into the message. Nor does the proportion of AI-generated words necessarily show how much the sender contributed to the final message. AI ticks should therefore not be presented as certificates of authorship or authenticity. They can report either AI assistance recorded by the platform or assistance voluntarily declared by the user. In unclear cases, the system could provide a simple description of the process without trying to measure the sender's authentic contribution, for instance, for assistance recorded by the platform, by displaying "AI-generated and then edited by the sender". Recent evidence on apology messages supports this distinction: disclosure of minor AI rewriting increased perceived authenticity and willingness to forgive, whereas disclosure of a largely AI-generated draft reduced both (Fan et al., 2026).

In light of these issues, we do not take a definitive stance on how this tool should operate. AI ticks should be considered here as a design framework requiring empirical validation, rather than a solution whose effectiveness is already established. We acknowledge that empirical work could help clarify the relational dynamics that emerge in these different cases and whether AI ticks could prove beneficial.

Nevertheless, we believe these insights move in the right direction and may provide useful cues for assessing the authenticity of interaction. In this sense, the transparency requirement functions not as a guarantee of authenticity but as a tool that enables individuals within a relationship to maintain greater control over their mutual expectations and, consequently, over authenticity-based reciprocal obligations.

4.2. Risks, Power Asymmetries, and Communicative Inequality

Disclosure is not ethically neutral. It can create problems of its own. Recipients may see even limited AI assistance as a sign of insincerity or lack of personal engagement, regardless of the sender's intentions or actual contribution. Requests for transparency may also arise in relationships marked by unequal power. Employers, supervisors, or intimate partners may pressure others to disclose their use of AI even when there is no legitimate expectation of authenticity. Accordingly, a preference to "want to know" should be treated as a request that the parties must negotiate. It should not give recipients an automatic right to disclosure or the technical ability to block or penalize messages.

The costs of disclosure may also fall more heavily on some users than on others. Non-native speakers may rely on translation or stylistic assistance to communicate on more equal terms, while persons with disabilities may use AI as an accessibility tool. For these users, AI may support agency rather than replace it. Treating such uses as evidence of reduced authorship could therefore create unfair stigma. Requiring users to explain why they need linguistic assistance might also reveal information about their disability, literacy, or linguistic background that they have no duty to disclose. Granular ticks should thus indicate what the AI did (for instance, translating, correcting language, reformulating text, or generating content) without requiring users to explain why they used it.

These risks do not undermine the case for AI ticks; they only place limits on how such ticks should be designed. Disclosure should be specific to the context and limited to the relevant interaction. Users should remain free to change their disclosure settings. Information about AI use should not become a permanent feature of a user's profile or a basis for penalties or rankings. Interfaces should also avoid implying that linguistic assistance is necessarily less authentic than unassisted expression. Empirical studies will be needed to examine how different labels are understood across relationship types, linguistic and cultural settings, and different accessibility needs. AI ticks should neither decide which expectations are legitimate nor force users to meet them. Their role is more limited: to provide the kind of frictional cue described by Seberger and colleagues (2026), making AI mediation visible while leaving the parties to negotiate what it means for their relationship. Transparency neither replaces nor guarantees authenticity. Its role is to clarify how a message was produced and to preserve the space in which users can negotiate their expectations about authenticity.

5. Conclusion

The rapid integration of AI into everyday messaging systems is transforming the conditions under which interpersonal communication takes place. As AI-MC becomes more seamless and context-sensitive, it increasingly shapes not only the form of our messages but also the social expectations that govern how we relate to one another. This development heightens an already pressing normative concern: whether communicative acts produced with AI assistance can still count as authentic expressions of the agent.

Our analysis has shown that authenticity is not a feature that can be engineered directly into communication systems. It depends on the user's reflective endorsement, on the relational norms appropriate to the communicative context, and on socially embedded practices of meaning-making. These dimensions are irreducible to technical parameters and cannot be codified in algorithmic rules. As a result, attempts to "design for authenticity" misconstrue the nature of the concept. What designers *can* do is avoid undermining authenticity and support the conditions under which users can exercise it.

This is where transparency plays a substantive role. While transparency cannot guarantee authenticity, it can prevent misunderstandings about the origin of communicative content and enable users to manage relational expectations appropriately. Offering forms of disclosure – whether unilateral, reciprocal, or negotiated – allows users to situate AI assistance within the moral and social landscape of their interactions. Such mechanisms do not dictate what authenticity requires, but they make its pursuit possible under conditions of AI support.

The broader challenge for AI-MC, therefore, is not to replicate or enforce authentic expression, but to preserve the normative space in which authentic agency can still operate. Systems that enable user control, calibrated disclosure, and contextual awareness might be better positioned to sustain trust and relational integrity than systems that obscure the distribution of agency behind the scenes. As AI becomes an increasingly common presence in personal communication, the task ahead is to design messaging environments that respect authenticity, transparency, and responsibility, values that remain central to the ethical fabric of human interaction.

Despite recent progress, AI-mediated communication remains a field marked by unresolved theoretical and technical questions. We still lack a clear account of how users' preferences, relational expectations, and contextual cues should interact with algorithmic systems that increasingly infer, predict, or even anticipate communicative intentions. Technically, the mechanisms through which messaging platforms might implement meaningful yet non-intrusive forms of disclosure remain underexplored, particularly as AI models become more deeply embedded and personalized. Moreover, the normative implications of AI systems that can learn from, or shape, relational dynamics remain largely unexplored. Addressing these questions will require sustained interdisciplinary collaboration through a research agenda encompassing, among others, three areas: empirical research to understand users' expectations and vulnerabilities, examine the effectiveness of AI ticks through experimental comparisons of default settings and binary versus granular disclosure, investigate recipients' interpretations across different relationship types, and explore

cultural and linguistic variation in expectations of authenticity; technical research to develop transparent and controllable systems; and philosophical inquiry to refine the concepts that should guide responsible design. Much remains to be done, but recognizing the complexity of these challenges is a necessary step toward developing communication technologies that enhance, rather than erode, the human practices they mediate.

Funding: This work was partially supported by a grant. Details have been omitted for double-blind peer review.

References

.

Bar, F., Weber, M. S., & Pisani, F. (2016). Mobile technology appropriation in a distant mirror: Baroquization, creolization, and cannibalism. *New Media & Society, 18*(4), 617–636. https://doi.org/10.1177/1461444816629474

Basoah, J., Cunningham, J. L., Adams, E., Bose, A., Jain, A., Yadav, K., Yang, Z., Reinecke, K., & Rosner, D. (2025). Should AI mimic people? Understanding AI-supported writing technology among Black users. *Proceedings of the ACM on Human-Computer Interaction, 9*(7), 1–51. https://doi.org/10.1145/3757423

Battisti, D. (2025a). Second-person authenticity and the mediating role of AI: A moral challenge for human-to-human relationships. *Philosophy & Technology, 38*(1), 28.

Battisti, D. (2025b). Co-authoring apologies with ChatGPT can support relationships but not second-person authenticity: A reply to Lavazza. *Philosophy & Technology, 38*(2), 72.

van Berkel, N., Skov, M. B., & Kjeldskov, J. (2021). Human–AI interaction: Intermittent, continuous, and proactive. *Interactions, 28*(6), 67–71.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT (NBER Working Paper No. w34255). National Bureau of Economic Research. http://www.nber.org/papers/w34255

Chen, M.X., Lee, B.N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A.M., Chen, Z., Sohn, T., Wu, Y.: Gmail Smart Compose: Real-Time Assisted Writing. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2287–2295. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330723 .

Craft, S., & Vos, T. P. (2021). The ethics of transparency. In L. T. Price, K. Sanders, & W. N. Wyatt (Eds.), *The Routledge companion to journalism ethics* (pp. 175–183). Routledge.

Fan, G., Liu, D., & Pan, L. (2026). Is it still you? Attributing authorship and authenticity in AI-assisted romantic communication. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Article 1294, pp. 1–24). ACM. https://doi.org/10.1145/3772318.3790762

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.

Frankfurt, H. G. (1988). *The importance of what we care about: Philosophical essays*. Cambridge University Press.

Fröding, B., & Peterson, M. (2025). Is an eloquent love letter written by AI better than a clunky one written by your partner? *Philosophy & Technology, 38*(2), 70.

Fu, Y., Foell, S., Xu, X., & Hiniker, A. (2024). From text to self: Users' perception of AI-mediated communication tools on interpersonal communication and self. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). https://doi.org/10.1145/3613904.3641955

Glikson, E., & Asscher, O. (2023). AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. *Computers in Human Behavior, 140*, 107592.

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100.

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J. T., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports, 13*(1), 5487.

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3290605.3300469

Komninos, A. (2024). Agency and morality as part of text entry AI assistant personas. In *Designing AI personalities: Enhancing human-agent interaction through thoughtful persona design* [Workshop at the ACM International Conference on Mobile and Ubiquitous Multimedia (MUM 2024)], Stockholm, Sweden. https://arxiv.org/abs/2412.08360

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.

Lavazza, A. (2025). AI, authenticity, and the chatbot between us: Commentary on Battisti. *Philosophy & Technology, 38*(2), 49.

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–22). https://doi.org/10.1145/3706598.3713778

Lee, M., Gero, K. I., Chung, J. J. Y., Shum, S. B., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E. A., August, T., Bhat, A., Choksi, M. Z., Dutta, S., Guo, J. L. C., Hoque, M. N., Kim, Y., Knight, S., Neshaei, S. P., … Siangliulue, P. (2024). A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–35). https://doi.org/10.1145/3613904.3642697

Lipovetsky, G. (2021). *Le sacre de l'authenticité*. Gallimard.

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Ngaihlian, D. (2025). Machine learning algorithms: Simulating intentionality in artificial intelligence. SSRN. https://doi.org/10.2139/ssrn.5271061

Redaelli, R. (2025). Intentionality gap and preter-intentionality in generative artificial intelligence. *AI & Society, 40*(4), 2525–2532.

RIZZ - Apps on Google Play (n.d.), https://play.google.com/store/apps/details?id=com.rizzlabs.rizz&hl=en_GB, last accessed 2026/09/04.

Robbins, S., & Blundell, I. (2026). Losing our voice? Generative AI and the degradation of human expression. *Minds and Machines, 36*(1), 2.

Seberger, J. S., Datta Gupta, S., Razi, A., & Rezapour, S. (2026). Communicative surrogates and the ethics of LLM-mediated communication. *AI & Society*. https://doi.org/10.1007/s00146-026-03096-5

Shneiderman, B. (2021). Human-centered AI: A new synthesis. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (pp. 3–8). Springer.

Taplio - LinkedIn Post Generator: Write Posts with AI (n.d.), https://taplio.com/create-linkedin-post-with-ai, last accessed 2026/09/04.

Taylor, C. (1992). *The ethics of authenticity*. Harvard University Press.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112.

Varga, S., & Guignon, C. (2023). Authenticity. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.). https://plato.stanford.edu/archives/sum2023/entries/authenticity

Velleman, J. D. (2005). The self as narrator. In J. Christman & J. Anderson (Eds.), *Autonomy and the challenges to liberalism: New essays* (pp. 56–76). Cambridge University Press.

Viber - AI on Viber (n.d.), https://help.viber.com/hc/en-us/articles/31738053588509-AI-on-Viber, last accessed 2026/09/04.

WhatsApp - Get the Tone of Your Message Right with Private Writing Help (n.d.), https://blog.whatsapp.com/get-the-tone-of-your-message-right-with-private-writing-help, last accessed 2026/09/04.

Will - manages your social media (n.d.), https://www.heywill.ai, last accessed 2026/09/04.

YourMove AI - AI for Dating Apps: Profiles, Photos & Replies (n.d.), https://yourmove.ai/, last accessed 2026/09/04.

Zhu, J., & Molnar, A. (2025). Blissful (A) ignorance: People form overly positive impressions of others based on their written messages, despite wide-scale adoption of generative AI. *arXiv* preprint arXiv:2501.15678.

AI disclosure: As non-native English speakers, the authors used ChatGPT and Claude exclusively for language support. All ideas, arguments, and substantive content remain entirely the authors' own.

Declarations: One of the authors received grant funding during the preparation of this manuscript. This information has been disclosed solely for transparency and will be included only after the completion of the peer review process.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Who Am I Really Talking To? Authenticity, Transparency, and the Case for "AI Ticks" in AI-Mediated Communication

Supplementary Materials - Appendices

A. Walkthrough example of AA-GP assistant use.

This appendix illustrates the cognitive effort and friction in using an AA-GP assistant to generate a contextually relevant and coherent reply to an incoming message. We take here the case of Miranda, an employee at an organisation. She receives an SMS from her supervisor, Chris:

> *"Hey Miranda, just wondering if that report is still on track for today? I need to manage expectations from my favourite girl on my end! Thanks!"*

The comment shifts responsibility onto her and includes a gendered expression that undermines her professionalism. Miranda feels upset but feels compelled to respond quickly and diplomatically. She decides it might be wise to consult an AI assistant (e.g. ChatGPT) to help craft a dispassionate response. To invoke the help of the assistant, she must switch apps, paste the message, and provide relevant contextual instructions, for example:

> *"I received the following message from Chris, my boss. The report is due next Friday. Write a short, polite, and professional reply explaining that I cannot complete it earlier than Monday. Do not exceed 140 characters."*

Even with this level of detail, the output may be unsuitable. If Miranda forgets to specify the character limit, the assistant may generate a long, email-like paragraph. If she omits the relational context, the AI may adopt an overly casual tone. If she does not forbid abbreviations, it may produce SMS shorthand. Achieving the desired result often requires multiple iterations: correcting the AI's assumptions, restating constraints, or editing the generated text. Each refinement requires a fresh round of input, gradually nudging the assistant toward a message compatible with Miranda's intentions. The cognitive load and friction in generating the desired output through an intermittent interaction process that starts with a context-rich but imperfect prompt, is illustrated in Figure 1.

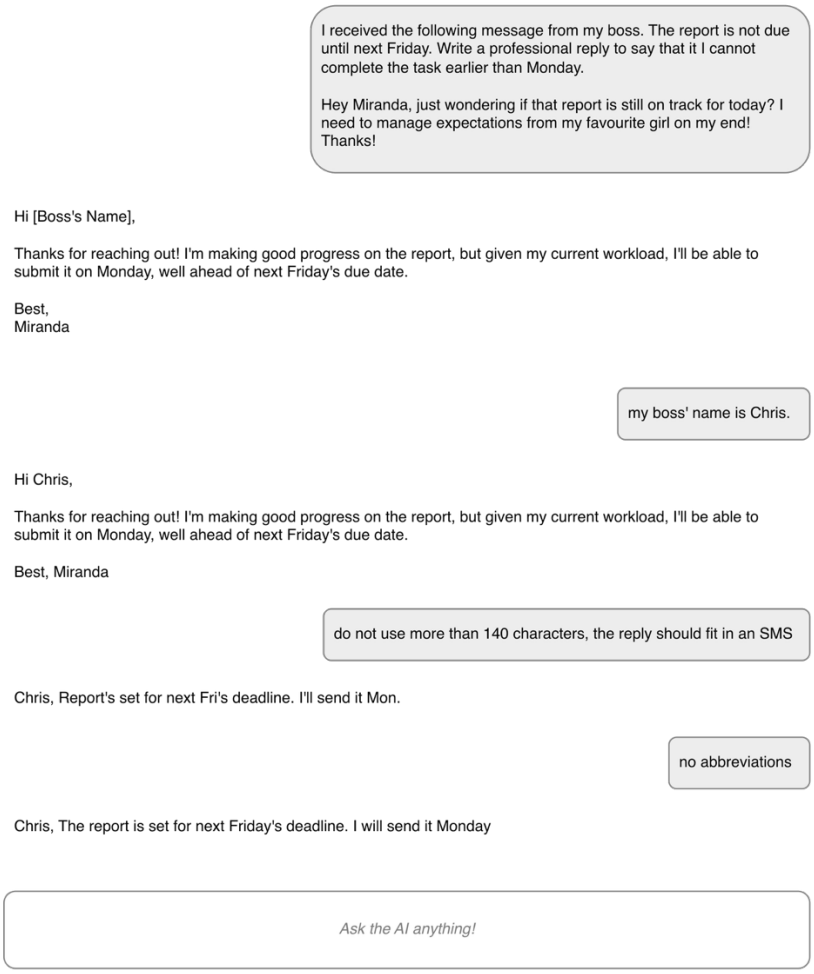


Fig. 1. *Imperfect instructions can lead to unsuitable output when using GPAs for communication assistance. Achieving the desired reply often requires iterating prompts, copying suggested answers, and editing them before sending. In this example, the users provides an initial prompt offering partial context to the AI assistant, and an example output of the assistant, followed by additional instructions to "fine tune" the output as originally intended by the user, in a dialectic process of intermittent interaction.*

B. Operating principles of mobile virtual keyboards

In this appendix we describe in simple terms the fundamentals of smartphone operating systems and how mobile virtual keyboards are "connected" to applications, so that they can provide input (see technical documentation for Android (n.d.) or iOS (n.d.)). This technical understanding is necessary to understand the capabilities and limitations of a keyboard-integrated AA-CS.

The virtual keyboard is simply another app installed on the device, for all intents and purposes much the same as any other user-installed app (which is why a user can download and install multiple virtual keyboards on their device, as they might do, for example, with camera apps). Smartphone operating systems run all applications in a protected execution environment, meaning they prohibit applications from directly accessing the memory space and data held by other applications. This is crucial to enforce safety, security and robustness of the entire system, protecting its stability and sensitive user data from malicious (or carelessly programmed) applications (Niemi, 2024). Therefore, for the keyboard app to be able to send data (keystrokes) to a messaging app or retrieve data from it (e.g. the contents of a text field, or cursor position), the operating system provides a "bridge" between an application's focused input field (e.g. a textbox) and the smartphone virtual keyboard (e.g. see Diao et al. 2018). This allows the two apps to securely exchange data only for the purpose of creating a new message, while protecting the messaging app's other data (e.g. conversation histories, contact details etc.) from access. The principles are illustrated in Figure 2.

As a result of the internal workings of the operating system and how input is enabled from the keyboard app to all other apps that invoke its use, any AI agent integrated within the keyboard is only able to perform de-contextualised operations (i.e. without knowledge of previous communications), and only on manually drafted texts. These AI agents cannot generate a reply from scratch as a response e.g. to the last incoming message (since the operating system does not offer a "bridge" to the application data, i.e. the last incoming message, which would hold the required contextual information to allow this generation).

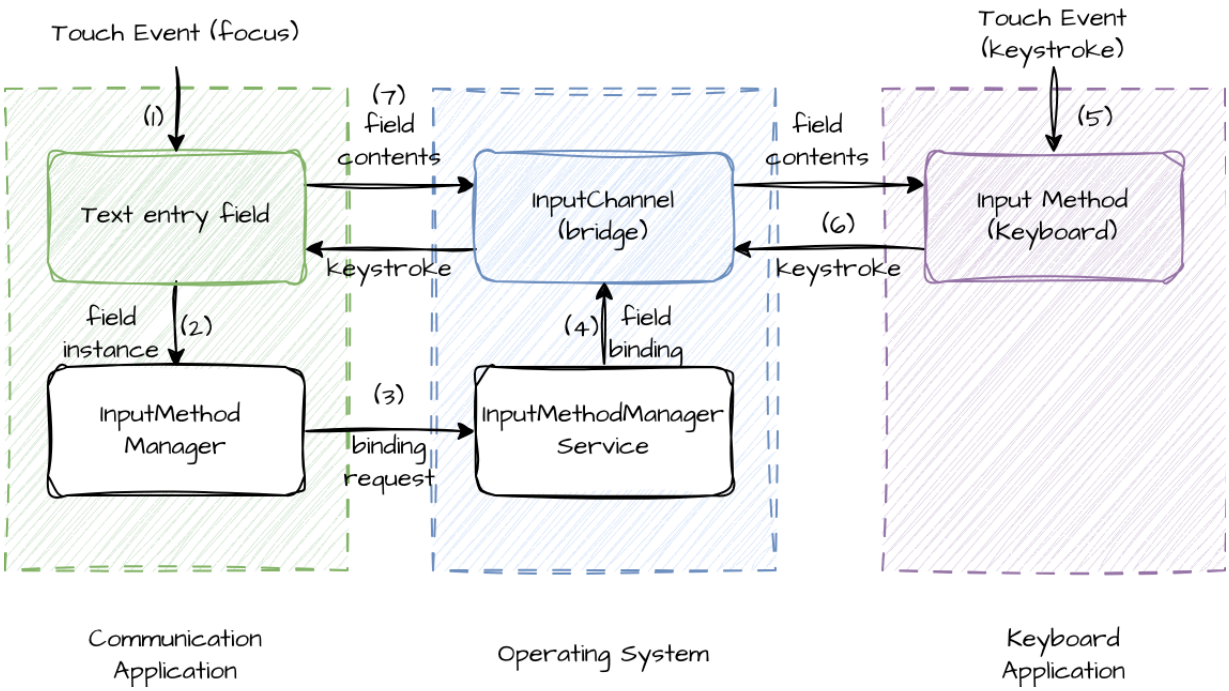


Figure 2. *Conceptual illustration of the "bridge" mechanism allowing data exchange between apps and keyboards on a typical smartphone operating system. Once a user places focus on a text entry field (1), the application requests to the operating system to create a binding between that specific text field (2) and the keyboard (3). The operating system grants the request and binds the text find with the user's default keyboard (4). This creates the data exchange "bridge" between the keyboard, through which users touch events on the keyboard area (5) are translated to keystrokes and sent to the text field (6). The keyboard can also retrieve the bound text field's contents (7) but does not have any other way to access other application data.*

C. Description of the AI-MC core attributes

Integration: Relates to the implementation of source code invoking the AI agent, in relation to the communication application were its assistance is required. Note here that the actual AI agent itself may be part of the communication app, or integrated within the operating system (available to all apps), or even hosted externally the user's device (on-cloud). Integration refers only to the source code invoking the AI, regardless of its actual location. External means the source code is outside the application, i.e. the AI is invoked from a third-party app or website. In-app means the source code is inside the application, thus the user does not have to switch application contexts to invoke the AI. Keyboard means the source code is integrated within the virtual keyboard. While technically this is also an external integration, the user perceives it as integrated inside the app, since the application context is not switched.

Function: This attribute refers four possible uses of the AI. Restyling means re-writing some text originally entered manually by the user, into a different linguistic tone (e.g. "cu @ 4" converted to a more professional "see you at 4pm".). Generation means creating an entire new message, typically as a response to an incoming message. Correction means to use the AI to correct spelling or syntax mistakes in a manually drafted message (e.g. "I find acomodation" to "I will find accommodation"). Finally, Completion refers to continuing a user's unfinished draft message (e.g. "hey Maria, sor|" completed to "hey Maria, sor|*ry I can't meet up tonight* ").

Contextual Access: This concerns the level of contextual information that the AI can automatically absorb from within the user's conversation environment. In broad terms, AA type applications largely depend on users to provide adequate contextual information by typing/pasting the necessary information as part of their prompt or dialogical process, or by providing screenshots of messaging history (Lavender), therefore have no contextual access. A low level of access to contextual information can be automatically obtained by keyboard-integrated AI, constituent of the message a user has typed for restyling (e.g. GBoard). On the other hand, AS integrations offer differing levels of automatic context absorption. Services like WhatsApp obtain low (limited) context in the same manner as AA-CS keyboard integrations, while Viber has moderate level access due to partial access to conversation history (up to 10 messages). SmartCompose has low to moderate access, as the context automatically retrieved depends on the current message being typed, or the immediate message thread context. Services like Taplio, Will and GMelius have high levels of access, since they allow users to import their entire messaging history to automatically infer context and personalise outputs.

User Control: This refers to the users' ability to parametrise and control AI outputs through prompting instructions. AA-GP applications offer a high level of user control and can be used to both generate and edit messages at will, through appropriate prompting. AA-CS integrations like keyboards, as well as some AS-CS like WhatsApp, only allow users a low level of control, as they can only select from a range of pre-generated options under undisclosed prompting instructions, which can of course be manually edited before sending. Finally, other AS-CS services like Viber or SmartCompose do not offer any user control, apart from accepting generated options.

Prompting Effort: This attribute is a direct result of the combination of contextual access and user control. No prompting effort is required in some AS-CS implementations like SmartCompose (Chen et al., 2019), Rizz (n.d.), or Viber (n.d.), as they do not allow users to directly affect how output is generated. AA-CS keyboard integrations and

WhatsApp (AS-CS) allow a limited way of prompting by affording users control over the linguistic style. Users simply select their preference to get a generated message, which internally affects the prompt, but in a manner that is completely abstracted away from the user. The most prompting effort is required by AA-GP implementations (e.g. Messenger).

Personalisation: Related to contextual access, personalisation refers to the AI's ability to learn from users' past behaviour and generate authentic-looking messages. AA-GP implementations offer no personalisation. To the best of our knowledge, AA-CS keyboard integrations, and AS-CS integrations like WhatsApp and Rizz do not offer personalisation either. Moderate levels of personalisation can be found in AS-CS integrations like SmartCompose since it learns from users' typing habits (but not message history). High levels of personalisation can be found in AS-CS implementations like Taplio (n.d.) or Will (n.d.), as they are able to ingest large volumes of communication history and learn users' writing styles.

AI Contribution: With reference to the amount of text that can be attributed to AI instead of the users, implementations can range from low (e.g. spelling corrections only) to high (entire messages generated from scratch). Broadly, due to their flexibility, AA-GP implementations afford the full range of contribution levels. AA-CS keyboard integrations and AS-CS WhatsApp can range from low-moderate as they only spell-check or modify original text entered by the user. AS-CS integrations like SmartCompose offer low levels of AI contribution as they only complement text written by the user, which may contain AI suggested words of course. Finally AS-CS implementations like Viber, Rizz, Taplio and Will, offer high levels of AI Contribution, as they can generate entire messages without user intervention.

References

Android - Create an input method | Views (n.d.), https://developer.android.com/develop/ui/views/touch-and-input/creating-input-method, last accessed 2026/09/04.

Chen, M.X., Lee, B.N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A.M., Chen, Z., Sohn, T., Wu, Y.: Gmail Smart Compose: Real-Time Assisted Writing. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2287–2295. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330723.

iOS - Creating a custom keyboard (n.d.), https://developer.apple.com/documentation/uikit/creating-a-custom-keyboard, last accessed 2026/09/04.

RIZZ - Apps on Google Play (n.d.), https://play.google.com/store/apps/details?id=com.rizzlabs.rizz&hl=en_GB, last accessed 2026/09/04.

Niemi, A.: Survey of Real-World Process Sandboxing. In: 2024 35th Conference of Open Innovations Association (FRUCT). pp. 520–531 (2024). https://doi.org/10.23919/FRUCT61870.2024.10516417.

Diao, W., Liu, R., Liu, X., Zhou, Z., Li, Z., Zhang, K.: Accessing mobile user's privacy based on IME personalization: Understanding and practical attacks12. Journal of Computer Security. 26, 283–309 (2018). https://doi.org/10.3233/JCS-16909.

Viber - AI on Viber (n.d.), https://help.viber.com/hc/en-us/articles/31738053588509-AI-on-Viber, last accessed 2026/09/04.

Taplio - LinkedIn Post Generator (n.d.): Write Posts with AI, https://taplio.com/create-linkedin-post-with-ai, last accessed 2026/09/04.

Will - manages your social media (n.d.), https://www.heywill.ai, last accessed 2026/09/04.