International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Human-Centered AI Language Technology (HCAILT): An Empathetic Design Framework for Reliable, Safe and Trustworthy Multilingual Communication

Vicent Briva-Iglesias & Sharon O'Brien

To cite this article: Vicent Briva-Iglesias & Sharon O'Brien (06 Feb 2026): Human-Centered AI Language Technology (HCAILT): An Empathetic Design Framework for Reliable, Safe and Trustworthy Multilingual Communication, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2622588

To link to this article: https://doi.org/10.1080/10447318.2026.2622588

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2622588

Taylor & Francis
Taylor & Francis Group

Human-Centered AI Language Technology (HCAILT): An Empathetic Design Framework for Reliable, Safe and Trustworthy Multilingual Communication

Vicent Briva-Iglesias and Sharon O'Brien

Dublin City University, Dublin 9, Ireland

ABSTRACT

Language technologies are increasingly ubiquitous and now translate emergency bulletins, draft clinical notes and mediate everyday conversations, yet their impressive fluency can be misleading–masking limited reliability, unpredictable errors and uneven performance across different user groups and languages. Building on Shneiderman's human-centered AI (HCAI) paradigm, this article introduces the Human-Centered AI Language-Technology (HCAILT) model, a domain-specific framework that binds reliability, safety culture and trustworthiness to the full language-technology pipeline. HCAILT couples technical guardrails (such as retrieval-augmented generation and quality estimation) with organizational practices (like bias audits and incident-report loops), together with user-facing features that maintain meaningful human control. Two blueprint use cases–in multilingual healthcare and crisis communication–illustrate how the HCAILT model guides system architecture, deployment practices and evaluation. A demo system demonstrates immediate feasibility on public large language models. By translating HCAI principles into actionable design levers, HCAILT provides scholars, developers and policymakers with a pragmatic path from ethical aspiration to deployable practice. The paper concludes with a research agenda for empirical validation in real-world settings and invites multidisciplinary collaboration to ensure that next-generation language technologies are not merely powerful, but demonstrably reliable, safe and worthy of public trust.

KEYWORDS

Human-centered artificial intelligence; multilingual digital communication; artificial intelligence; language technologies; human-computer interaction

1. Introduction

Languages, with all of their richness and intrigue and their encoding of culture and knowledge, remain both a pathway to mutual understanding and, paradoxically, a barrier to human-to-human communication and cooperation, especially in high-stakes contexts such as medical emergencies or disaster response.

Over recent decades, language technologies–systems that "enable machines not only to read, analyse, process and generate human language, but also, thanks to recent scientific advancements, to bridge the divide between human communication and machine understanding"[1] – have progressively expanded from translation-memory databases used by professional translators since the 1990s to today's generative artificial intelligence (AI) models capable of speech-to-speech translation, automatic text translation (or machine translation (MT)), or summarization, among other language generation tasks (Briva-Iglesias, 2023; Brown et al., 2020). Yet nowadays, translators are no longer the only users of language technologies, and the scale and speed of generative AI adoption outpace critical scrutiny: hallucinations, errors and latent biases continue to surface, often where the social cost of failure is highest (Weidinger et al., 2022).

To investigate these challenges, we adopt Shneiderman's Human-Centered AI (HCAI) paradigm, which foregrounds reliable systems, a culture of safety, and trustworthiness through empathetic design and meaningful human control (Shneiderman, 2022). While HCAI has gained traction in

CONTACT Vicent Briva-Iglesias vicent.brivaiglesias@dcu.ie Dublin City University, Collins Ave Ext, Whitehall, Dublin 9, Ireland

human–computer interaction, a well-established field within computing (Shneiderman, 2022), its application to language technology applications remains under-theorised and under-developed (Briva-Iglesias, 2024; O'Brien, 2024). Generic human-centered AI principles –such as the AI4People framework (Floridi et al., 2018) or the 2024 EU AI Act's risk tiers (EU, 2024)- mention language technology systems only in passing, although empirical work has already shown, for instance, that faulty language technologies can jeopardize multilingual communication or amplify misinformation (Yao et al., 2024). A HCAI consideration of language technology is therefore overdue.

We answer that call by proposing the Human-Centered AI Language Technologies (HCAILT) model, a framework that prioritizes the goals of reliability, safety and trustworthiness proposed by Shneiderman (2022) across the entire language-technology pipeline. HCAILT recognizes two fundamental drivers: (i) augmenting human cognition by reducing cognitive load and enabling accurate decision-making in multilingual settings; and (ii) augmenting information dissemination by delivering rapid, accurate, context-appropriate communication across linguistic boundaries in both routine and life-critical scenarios.

This article makes three interlinked contributions. First, conceptually, by formalizing the HCAILT model and translating the HCAI ideals into concrete design levers tailored to AI-powered language technology systems. Second, empirically, by illustrating the HCAILT model through two real-world use cases: multilingual healthcare communication and crisis communication, where language barriers carry life-or-death consequences (Villarreal et al., 2025). Finally, practically, by presenting a Vercel demo that exemplifies HCAILT's guardrails and provides a blueprint for researchers, developers and regulators. These contributions are guided by two overarching research questions. Within the HCAI framework:

- RQ1. How can reliability, safety and trustworthiness be concretely operationalized in language-technology workflows?
- RQ2. What constitutes a robust framework for the evaluation of reliability, safety, and trustworthiness in language-technology workflows?

We first introduce the main components of Shneiderman's HCAI paradigm in Section 2. Section 3 applies the HCAI paradigm to the domain of language technology, introducing the proposed HCAILT framework. Section 4 provides two examples of real-world use cases in the healthcare sector and in crisis response. Section 5 introduces a demo system that applies the principles of the HCAILT framework and a preliminary approach to evaluation. We conclude by offering some reflections on limitations, challenges and ways forward.

2. Shneiderman's HCAI framework

HCAI is described by Shneiderman (2022, p. 3) as an expansion of an algorithm-focused view of AI to a human-centered perspective that will shape the future of technology to better serve human needs. Process and product are two key aspects. Process involves user observation and stakeholder engagement to evaluate human performance in use of systems that employ AI and machine learning. Product, on the other hand, is focused on HCAI systems that are designed to "augment, empower, and enhance human performance" (ibid: 9) while emphasizing human control. Shneiderman's HCAI framework aspires to "high levels of human control AND high levels of automation" (ibid: 9) where previously one was considered to rule out the other.

The HCAI framework requires the implementation of governance structures to achieve three goals: (1) Reliable systems; (2) a Safety Culture; and (3) Trustworthiness. A requirement for these three goals is an empathetic design philosophy. As Shneiderman writes: "Empathy enables designers to be sensitive to the confusion and frustration that users might have and the dangers to people when AI systems fail, especially in consequential and life-critical applications" (Shneiderman, 2022, p. 20). For instance, in the translation industry –one of the early areas of language technology adoption– it is no overstatement to suggest that empathy has typically not been shown to professional translators by MT software developers or researchers. The primary goal, both in professional and research settings, was to drive algorithmic advances to create more and better translation output, faster and cheaper, with little consideration of the cognitive effort involved in fixing the output (called post-editing) by professionals and even less

given to the impact on their professional wellbeing (Baumgarten & Bourgadel, 2024; Moorkens, 2024). Somewhat ironically, to reach these goals, developers of MT systems used parallel translation data created by professional translators, but without acknowledgement of this reuse and certainly with no reimbursement. This lack of empathy is also evidenced by the common term "human-in-the-loop", typically offered as a consolation by developers to the professional translation community, as a grudging acceptance of the need for human control in a process that would ideally–from the developer's perspective–be seamlessly automatic (Shneiderman, 2020). Below we detail the three goals of the HCAI paradigm and, in Section 4, we analyse how the use of AI-driven language technologies by general users can be fit into these proposed governance structures and how empathy can be factored in beyond a human-in-the-loop perspective, especially when there may be consequential and life-critical applications.

2.1. Reliable systems

In Shneiderman's HCAI Framework, reliable systems "produce expected responses when needed" (Shneiderman, 2022, p. 53). A number of factors control such reliable responses, the most relevant of which for language technologies general use are audit trails and analysis tools, and verification and bias testing to enhance fairness. Free to use, generic language technology tools such as Google Translate or ChatGPT, among others, were no doubt tested in those company's labs before release. However, one recurrent reliability concern is that, when given the same input in different instances, these systems do not necessarily produce the same output and may even return divergent, contradictory or clearly hallucinated responses across interactions (Ahmad et al., 2023; Asgari et al., 2025). For example, a public-health advisory during a pandemic might in one interaction be translated or summarized with the correct isolation period and dosage, and in another interaction with the same input text be rendered differently, producing conflicting guidance with serious implications for infection control and medication safety, even though both outputs appear fluent and well-formed.

In addition, "testing," in the case of AI-powered language technologies output, is notoriously problematic, especially if it is carried out by system developers instead of actual end users. For instance, in MT, the approach to testing has been to use the concept of a "gold standard" sentence and to compare the system's output for similarity to that sample sentence (Kocmi et al., 2021). While this is one form of validation it does not consider the fact that there is no such thing as one agreed translation for any one sentence, that meaning is communicated also at a textual level and is not restricted to sentence level, is context-dependent, and the scores provided are meaningless to the general user (Freitag et al., 2021; Kenny, 2022). Audit trails and analysis tools are in use when systems are being developed and improved. However, mechanisms for auditing once a system has been released into the "wild" are limited. Professional translators will seek out the worst possible mistranslation and profile that publicly to demonstrate that systems are faulty (usually without overtly recognizing that human translators, even professionals, also make mistakes). However, this could hardly be considered an audit trail or analysis.

2.2. Safety culture

The HCAI framework proposes the building of a safety culture through business management practices. Shneiderman proposes five mechanisms that can help establish a safety culture: (1) leadership commitment to safety, (2) hiring and training oriented to safety, (3) extensive reporting of failures, (4) internal review boards and (5) alignment with industry practices. There is little evidence to suggest that these safety cultures have been considered and embedded while developing language technologies for general use. As mentioned above, the business model is more one of: who can produce the best system for specific languages to generate revenue. The fact that a speech-to-speech translation system could be used by a paramedic to communicate with a woman giving birth in an emergency, for example, or by an immigration officer deciding on the legitimacy of an asylum seeker's case, has been given no attention from a safety perspective. It should be acknowledged that the lack of evidence of safety measures does not necessarily mean that none have been considered or implemented. However, it is fair to say that narratives on safety considerations for the general use of AI-powered language technologies are not evident in the public domain. Mistranslations can have consequences, some much more serious than others.

A leadership commitment to safety would ensure that an AI-driven language technology system would not be used in circumstances where it should not be used, by people who do not understand that it can be faulty, and it would have auditing and reporting capabilities built in when it is released for general use. Short disclaimers at the bottom of a webpage hardly qualify as a way to ensure safe use of this technology.

2.3. Trustworthiness

Some research has been carried out on the topic of trust and language technologies. Focusing on general user usage, Rossetti et al. (2020) conducted a survey to understand the impact of MT and post-editing awareness on comprehension of and trust in messages disseminated to prepare the public for a weather-related crisis. All messages presented to participants were in fact machine translated, but participants were told that only some were machine translated. The authors found correlations between comprehensibility and trustworthiness, and identified other factors influencing these aspects, such as the clarity and soundness of the messages. The focus here was, however, on the MT outputs and not on the system or system developers per se. Gao et al. (2014) conducted an experiment involving a collaboration task between English and Mandarin speakers, hypothesizing that attributions about the source of errors affects collaboration experience. They found that beliefs about the presence of MT, which could also impact on trust, did in fact affect MT-mediated collaborations. Gao and colleagues go on to make some recommendations on designing for and with MT, first by making MT salient through interface design, providing explicit translation controls for senders and receivers of MT-mediated messages, providing notifications when the message may have been mistranslated, and by increasing perceived agency by allowing users to see the system as an active agent for communication. We revisit some of these ideas below by proposing a foundational theoretical model grounded in HCAI principles to guide the development, deployment, evaluation, and adoption of AI-powered language technologies.

3. The HCAI language technology (HCAILT) model

The primary objective of the proposed HCAILT model is to operationalize the principles of HCAI to enhance cross-lingual communication across diverse user groups. This model has been designed by having in mind the growing adoption of AI-driven language technologies by general users in various sectors, including but not limited to, healthcare (Briva-Iglesias & Peñuelas Gil, 2025), crisis communication (O'Brien, 2020), academia (Bowker & Buitrago-Ciro, 2019; Goulet et al., 2017) or public services (Vieira et al., 2023). The HCAILT model seeks to move beyond traditional algorithm-centric approaches, placing emphasis on human agency, cognitive augmentation, and empathetic, user-centered design. By operationalizing principles of reliability, safety, and trustworthiness, the model aims to maximize the societal benefits of AI-powered language technologies while minimizing potential risks.

3.1. Drivers of the HCAILT model

The proposed model (see Figure 1) is driven by two fundamental elements. The first driver of the HCAILT is augmenting human cognition. Human-centered AI-powered language technologies should reduce cognitive load, facilitate accurate decision-making, and empower users by enhancing their capabilities rather than merely replacing human work. This includes supporting real-time comprehension and interaction in multilingual contexts, thus improving decision-making quality and efficiency. For instance, AI-powered chatbots have been effectively employed in mental health interventions, providing initial screenings and preliminary psychological support, thus helping clinicians focus on more complex cases (Fitzpatrick et al., 2017). Similarly, AI-powered summarization and dialogue tools in healthcare have also demonstrated substantial potential by synthesizing large volumes of clinical data into concise, actionable insights, significantly reducing cognitive overload for healthcare providers (McDuff et al., 2025; Tu et al., 2025). Dorn (2025) highlights how these tools have begun reshaping patient-healthcare provider interactions by facilitating access to medical records and diagnostic information, enabling quicker, more informed clinical decision-making.

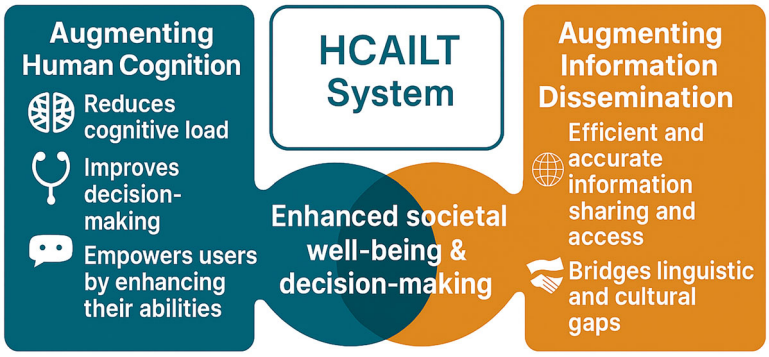

Figure 1. Drivers and impact of HCAILT systems.

The second driver of the HCAILT model is augmenting information dissemination. The model prioritizes efficient and accurate dissemination of information across linguistic and cultural barriers. It addresses both routine communications and critical, life-saving contexts by providing rapid, accessible, and context-appropriate multilingual information to diverse populations. Practical examples include the use of AI-generated multilingual advisories during public health emergencies, such as the COVID − 19 pandemic, where rapid and accurate communication of preventive measures, vaccination information, and travel advisories were critical (Xiao & Yu, 2025). Similarly, speech-to-speech translation tools integrated into clinical or emergency settings have improved communication between healthcare providers and patients in multilingual environments, reducing cognitive burdens associated with language barriers and enhancing clinical accuracy (Koutsouleris et al., 2022; Marais et al., 2020). Additionally, multilingual communication facilitated by AI has played a crucial role in humanitarian and disaster response, providing essential information rapidly to affected communities and emergency responders (Lewis, 2010), though its unmonitored use can also lead to inaccurate information (Pym et al., 2022).

By augmenting both human cognition and information dissemination, the HCAILT model ensures AI-driven language technologies significantly enhance human decision-making capacities, accessibility of critical information, and overall societal well-being while interacting with AI-powered language technologies.

3.2. Core components of the HCAILT model

The HCAILT model translates Shneiderman's HCAI model into concrete, domain-specific requirements for the design, deployment, and evaluation of language-technology systems. Each component is defined below together with the technical mechanisms, organizational practices, and user-facing features needed to satisfy it. Although conceptually distinct, the three components operate as interlocking layers: reliability provides the technical foundation; safety culture embeds those technical safeguards in accountable routines; and trustworthiness emerges when users can verify and understand system behaviour.

3.2.1. Reliability in the HCAILT model

Reliability in AI-powered language technologies refers to the consistent delivery of accurate, contextually appropriate, and timely communication outputs. In high-stakes contexts such as legal, healthcare and emergency management, unreliable systems can lead to severe consequences (Weidinger et al., 2021). Thus, reliability evaluation in HCAILT entails rigorous validation and verification processes, including accuracy evaluation tailored to specific domains, real-time latency (speed) measurements, and ongoing system performance monitoring.

Within the HCAILT framework, reliability can be enhanced through several mechanisms. One key strategy is the implementation of retrieval-augmented generation (RAG), which is a method where the AI system looks up information from trusted sources to improve its answers, rather than relying solely on its training data (Conia et al., 2024). This helps to minimize hallucinations (making up false information) and ensures that the output is grounded in verified facts (Li et al., 2022). Incorporating domain-specific datasets –e.g., using medical terminology officially validated by the World Health

Organisation– can further ensure output relevance and accuracy. Additionally, HCAILT systems may be configured to process only content from predefined domains, ensuring that general-purpose models are not misapplied to specialist communication. For example, an AI-powered language technology system designed for the legal domain should only provide an answer if the content of the interaction is within its specialization field. Otherwise, the system should suggest contacting a professional in that other domain.

Another important control measure for reliability is the temperature setting of Large Language Models (LLMs). Temperature refers to an LLM parameter usually ranging from 0 to 2 that governs its degree of randomness or creativity. A higher temperature (e.g., 1.5–2) results in more varied and creative outputs, whereas a lower temperature (e.g., 0–0.5) produces more deterministic and stable responses (Peeperkorn et al., 2024). For HCAILT applications, calibrating the temperature appropriately for the context is critical. In clinical or legal MT tasks, where precision and consistency are paramount, a low temperature may be necessary to avoid unintended variations. Conversely, slightly higher temperatures may be appropriate in plain-language rewriting tasks, where some degree of rephrasing is desirable. Determining the optimal temperature setting for each specific use case is therefore an essential part of system tuning and a practical measure for maintaining output reliability.

Finally, reliability can also be evaluated through controlled comparative assessments, where AI-generated translations are benchmarked against translations produced by professional domain experts. Studies such as McDuff et al. (2025) highlight the effectiveness of AI systems in healthcare scenarios, demonstrating that when rigorously tested, AI language technologies can reliably support clinical interactions and documentation, reducing errors in medical prescriptions, improving communication between patients and providers, and enhancing overall process quality.

3.2.2. Safety in the HCAILT model

The establishment of a robust safety culture is crucial within the HCAILT model, especially in sensitive domains. Safety encompasses proactive management practices, leadership commitments to ethical standards, comprehensive bias mitigation, and stringent privacy and security protocols. Empirical studies, such as those by Koutsouleris et al. (2022), illustrate that significant barriers persist regarding ethical AI practices, including ensuring transparency, interpretability, and fairness. Therefore, evaluating safety within HCAILT involves implementing mechanisms such as comprehensive bias audits, strict adherence to data privacy standards (e.g., GDPR and HIPAA compliance in healthcare), and ethical governance structures.

For instance, AI-driven mental health chatbots, such as Woebot, demonstrate the critical role of safety in the application of language technologies. Research by Fitzpatrick et al. (2017) and Yeh et al. (2025) underscores the importance of designing AI interfaces with robust ethical considerations, highlighting issues of language barriers and technical limitations that could inadvertently compromise patient safety by increasing anxiety or misinterpretation. Another critical safety measure involves enhancing user literacy regarding AI-powered language technologies, exemplified by the work on MT literacy (Bowker, 2020). Ensuring users understand the capabilities and limitations of these technologies significantly reduces risks associated with misuse or over-reliance. Educating users on the functionalities, limits, and appropriate contexts of use fosters realistic expectations and informed utilization, reinforcing that while AI systems augment human decision-making, ultimate responsibility for critical decisions remains with the human user (Doshi-Velez et al., 2019; Ojewale et al., 2025).

3.2.3. Trustworthiness in the HCAILT model

Trustworthiness captures the degree to which users justifiably rely on system outputs. It is not merely a psychological state but the outcome of verifiable safeguards that make system behaviour legible, contestable, and reversible. For language technologies, trust emerges from a combination of verifiable safeguards and legible interaction cues that help users assess residual risk, especially when perfect accuracy cannot be guaranteed.

At the heart of HCAILT's trust layer could lie an output-level Quality Estimation (QE) feature. This QE feature would assign a probability score indicating the likelihood that a given output contains an

error (Huang et al., 2023). QE does not claim to be infallible; rather, it offers a calibrated signal that higher attention should be paid to specific areas. Scores could also be mapped to a simple, colour-coded interface –e.g., green for high-confidence output, yellow for moderate confidence and red for low confidence– so that non-expert users can immediately visualize where additional scrutiny or human intervention is advisable. By foregrounding uncertainty, QE prevents the "automation bias" observed when users assume that AI-powered language technology outputs are either entirely correct or entirely wrong, thus significantly mitigating risks in critical contexts (Fomicheva et al., 2020).

To sustain long-term trust, HCAILT systems should foster independent certification and open evaluation. HCAILT systems should undergo third-party audits and publish benchmark artifacts –QE models, evaluation datasets and error logs– so that the wider community can replicate, critique and improve upon reported performance. Such transparency aligns with emerging regulatory proposals that treat high-risk language technologies in a manner similar to safety-critical medical devices (Chen et al., 2018). Taken together, probabilistic QE, colour-coded uncertainty visualization and independent control transform abstract principles of transparency and accountability into day-to-day interaction features that foster well-calibrated trust in a partnership in which humans remain decisively in command of AI-powered language technologies in multilingual communication.

4. Application: Real-world use cases

To illustrate the practical applicability and robustness of the HCAILT model, we reflect on how the model could be applied to two real-world use cases from different domains: healthcare communication and crisis communication. We first summarize domain-specific challenges from prior work and then show, via concrete scenarios, how HCAILT's components could structure system design and governance in these settings.

4.1. Use case 1: Healthcare communication

Multilingual healthcare environments, such as hospitals and mental health services, regularly encounter significant language barriers that complicate accurate diagnosis, informed consent, therapeutic rapport, and treatment adherence (Montalt-Resurrecció et al., 2024). Traditional solutions, primarily professionally trained and experienced human interpreters, often lack availability, particularly in emergency scenarios and for less commonly spoken languages, posing substantial risks to patient safety and healthcare effectiveness (Valero-Garcés, 2025). Additionally, healthcare organizations might not have adequate budgets for such services. In some cases, relatives of the patient may act as the ad hoc interpreter. These relatives may not have medical knowledge nor adequate language skills for interpreting specialized conversations, which poses a risk to the patient. Additionally, the use of family members, sometimes even minors, for mediating communication in these high-risk settings raises serious ethical questions (Antonini, 2016). AI-powered language technologies, informed by the principles of the HCAILT model, may allow for overcoming some of these linguistic and knowledge barriers effectively, though it also must be acknowledged that they are not without ethical concerns too. Practical examples of such technologies include speech-to-speech AI translation tools, enabling real-time, contextually precise conversations between patients and healthcare providers. These tools, if developed using appropriate and quality-controlled data, could maintain the integrity of medical terminology, significantly reducing potential errors. They can also be integrated with automatic transcription features, allowing healthcare providers and patients to review conversations afterwards and clarify potential misunderstandings (Wysocki et al., 2023). Another valuable technology involves AI-driven MT systems that are specifically fine-tuned to the medical domain and integrated into electronic health record systems. These ensure precise translations of medical documentation, patient histories, and treatment instructions, crucially minimizing errors and misunderstandings arising from cross-lingual communication (Rodríguez-Miret et al., 2024).

There are potentially multiple benefits for employing these AI-powered language technologies. Primarily, they significantly augment human cognition by reducing the cognitive load for healthcare providers and patients who may struggle with comprehension given limited language competence,

enabling accurate comprehension and more efficient decision-making processes. Studies such as Fitzpatrick et al. (2017) and Yeh et al. (2025) have demonstrated how AI-driven mental health chatbots effectively reduce cognitive strain by conducting initial mental health screenings, facilitating accurate diagnoses, and supporting therapeutic interactions. Such technologies also augment information dissemination, making essential healthcare information more accessible to linguistically diverse populations and thus potentially reducing disparities in healthcare access and outcomes (Montalt, 2021).

In operationalizing the HCAILT model for multilingual healthcare, several considerations are critical. Reliability mandates that AI-generated translations, speech-to-text or speech-to-speech transformations and interactions be consistently accurate. Reliability levels would depend on using appropriate, quality-checked data for training the system in the first instance, followed by rigorous validation and ongoing verification. Secondly, reliability would require timely delivery of content, ensuring healthcare professionals and patients can rely on the technology in real-time clinical scenarios which might require split-second decision making. Safety considerations encompass stringent privacy protections, bias mitigation, and ethical compliance, particularly in handling sensitive patient data and critical healthcare information (Koutsouleris et al., 2022). Trustworthy systems require the integration of robust guardrails such as RAG techniques and domain-specific medical glossaries to prevent inaccuracies or hallucinations, particularly with sensitive medical information. An option could be to state that the system is not able to offer help beyond a certain specialization (if designed and developed with that goal), fostering trust among healthcare professionals and patients.

By embedding these considerations into AI-powered language technologies, the HCAILT model ensures that multilingual communication in healthcare contexts is not only technically effective but ethically sound, ultimately promoting patient safety, healthcare provider confidence, and equitable healthcare delivery across diverse linguistic populations. It is worth stressing that these technologies will be easier to achieve with major languages, due to data availability, and quality and application will be most costly and less effective in minor languages, due to data sparsity (Briva-Iglesias, 2022).

4.2. Use case 2: Crisis communication

Effective multilingual communication is crucial during crises, such as pandemics, disasters, humanitarian emergencies, and geopolitical conflicts, where timely and accurate dissemination of information can be lifesaving (Federici & O'Brien, 2020; O'Brien & Federici, 2023). Miscommunication in such scenarios often leads to resource misallocation, increased panic, and preventable personal and material damages. Leveraging AI-powered language technologies within the HCAILT framework can significantly enhance communication efficiency and accuracy in these critical circumstances.

Potential technologies in this domain include AI-generated multilingual public advisories, which enable rapid and accurate dissemination of essential safety protocols, public health guidelines, and emergency updates across various languages. These advisories could use MT, automated summarization tools, and plain language generation to ensure information is accessible to diverse linguistic communities (Cadwell et al., 2024). Additionally, speech-to-speech MT technologies could be invaluable in multilingual disaster zones, facilitating real-time communication between first responders, field emergency workers, and affected populations who may not share a common language (Lewis, 2010). AI agents and context-aware systems can further assist emergency personnel by swiftly translating instructions, medical advice, and situational updates, even in low-connectivity or offline scenarios (Briva-Iglesias, 2025). The benefits of employing AI-powered language technologies in crisis communication can be substantial. Primarily, they augment human cognition by significantly reducing the cognitive burden on emergency responders. This enables faster and more accurate multilingual communication and support during high-stress situations. Such tools also enhance information dissemination by rapidly scaling emergency alerts, health advisories, and critical instructions, ensuring these are effectively delivered in multiple languages and diverse communication formats, including audio, text, and simplified language.

Implementing the HCAILT model within crisis communication contexts requires careful consideration of its core principles. Reliable systems are critical; therefore, AI models must be optimized for accuracy. As discussed in the use case scenario above, this would also involve training with appropriate,

quality-checked data and rigorous testing in advance. The model would also need to be optimized for low latency, even in environments with limited connectivity. Offline functionality is especially important for maintaining communication continuity in disrupted or remote regions, typical of disaster zones. Ensuring safety involves prioritizing human validation of high-stakes messages, rigorous model bias evaluation, and strict adherence to privacy protocols to protect sensitive information during emergencies (Xiao & Yu, 2025). Finally, trustworthiness demands stringent fact-checking mechanisms, such as RAG, to minimize the risk of AI-generated mis- and disinformation, particularly in sensitive public health advisories and instructions. Overall, by incorporating these considerations, AI-powered language technologies can effectively support critical communication needs during crises, ensuring reliability, safety, and trustworthiness, thereby enhancing emergency response and community resilience.

5. Demo system and potential evaluation

To demonstrate the practical applicability and feasibility of the HCAILT theoretical model, we developed an interactive demonstration system using Vercel, designed specifically for multilingual communication within healthcare scenarios. The demo exemplifies the principles of trustworthiness, reliability, and safety embedded in the HCAILT model, contrasting its user-centered enhancements with LLMs, the state-of-the-art AI-powered language technology at the time of writing. This demo can be accessed at the following URL: https://hcailt.vercel.app/.

5.1. Demonstration of the workflow

The demo has been designed by reflecting on a specific use case: the one of an Irish holidaymaker in Spain, who has limited Spanish proficiency, who experiences symptoms of chest pain. Having presented at an emergency department, this results in potential problems in healthcare provision because the patient cannot communicate properly with the healthcare providers and there is no professional interpreter available immediately. The patient is assessed and presented with some written information, in Spanish, on symptoms and medication. The interactive demo system consists of a structured multi-agent workflow (Briva-Iglesias, 2025), comprising three sequential AI agents, each performing specific tasks aligned with the HCAILT framework.

Figure 2 depicts the demo workflow. A user uploads a source document through the interface, triggering the Machine Translation Agent, which invokes a RAG module constrained to a vetted medical knowledge base. RAG grounding limits hallucinations and enforces domain scope; if the input falls outside the clinical domain, the agent halts and recommends human interpretation, satisfying the model's reliability guardrail.

The resulting draft translation is passed to the Quality Estimation Agent, which assigns a sentence-level error probability and aggregates these into an intuitive traffic-light score (green ≥ 0.80, amber

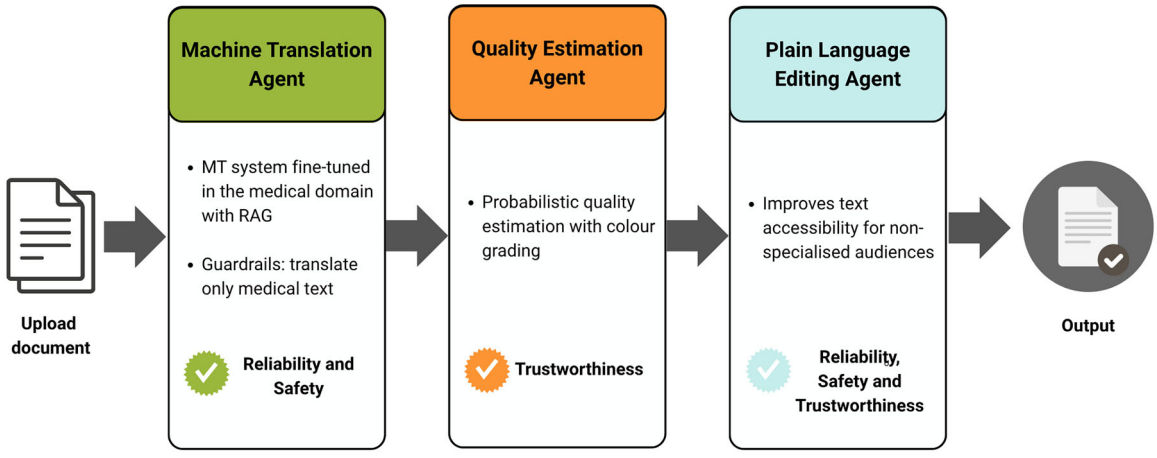

Figure 2. Step-by-step workflow of the demonstration system.

0.60–0.79, red $\leq$ 0.60). The score is presented both numerically and via colour overlays, enabling users to assess trust immediately and decide whether expert review is required.

Finally, the Plain Language Editing Agent simplifies the translated medical information into plain language suitable for patient comprehension. This step ensures accessibility and reduces cognitive load, allowing patients, particularly those with limited medical literacy or linguistic proficiency, to understand essential healthcare information clearly. The demo visually and interactively exemplifies the integration of these HCAILT principles, reinforcing human control and enhancing cognitive and informational accessibility in multilingual healthcare interactions.

5.2. Proposed evaluation strategy

While a comprehensive empirical evaluation remains beyond this manuscript's scope, we propose a structured, multi-dimensional approach for future evaluations of such systems to rigorously assess adherence to the HCAILT framework. This section addresses RQ2 at a conceptual level by outlining how reliability, safety and trustworthiness could be evaluated once HCAILT-compliant systems are deployed in practice.

Reliability evaluation would involve conducting translation accuracy checks by human experts (Läubli et al., 2020). Also, the HCAILT-enhanced system could be benchmarked against baseline outputs from traditional LLM systems, quantifying errors, terminology accuracy, and misinformation instances, demonstrating the benefits of introducing RAG and guardrails for a more reliable AI output. Additionally, system latency (number of tokens generated per second) would need to be assessed during real-time clinical scenarios to ensure usability under realistic healthcare time constraints.

Safety evaluation would focus on ensuring the system's adherence to patient data privacy regulations (e.g., GDPR, HIPAA), verifying its capability to consistently anonymize patient data, such as names, addresses, and dates, thus protecting patient confidentiality. Comprehensive bias audits would also be conducted, examining linguistic and cultural fairness to ensure equitable performance across various patient groups and linguistic backgrounds (Birhane, 2021).

Trustworthiness evaluation would involve recruiting bilingual health professionals or translators specializing in the health domain to assess the veracity of QE scores to further tune the QE system. Additionally, the use of colour coding to indicate different levels of probability in QE scores could be assessed for trustworthiness. Furthermore, usability and satisfaction surveys could be conducted with diverse user groups, including patients, healthcare providers, and language professionals, focusing on system transparency, ease-of-use, and perceived reliability. Such qualitative feedback would offer valuable insights into user trust dynamics and identify areas for further enhancement.

By incorporating these rigorous evaluation methods, future research will robustly validate the practical efficacy, ethical soundness, and overall alignment of the HCAILT model with real-world demands for reliable, safe, and trustworthy language technologies in both critical and routine contexts.

6. There are no pros without cons: Challenges of HCAILT systems and tools

Despite the evident benefits and potential improvements offered by the HCAILT model and related AI-powered language technologies, significant challenges persist in their development, deployment, and widespread adoption. These challenges must be acknowledged and addressed comprehensively to ensure these technologies' responsible and ethical use.

One of the primary challenges is ensuring consistent reliability and accuracy across different languages and domains. Data scarcity and imbalance for less commonly spoken languages significantly hinder the performance of AI models, creating disparities in accessibility and quality of communication (Pava et al., 2025). This issue is compounded by the inherent complexity and variability of human language, especially in context-specific scenarios such as medical, legal, and crisis situations.

Another crucial challenge involves user trust and technology transparency. Users' willingness to adopt and rely on AI-powered language technologies heavily depends on their confidence in the systems' outputs and the clarity of how these outputs are generated. Lack of transparency can lead to mistrust, reluctance, or misuse, undermining these technologies' effectiveness and potential benefits

(Doshi-Velez et al., 2019). Additionally, the rapid technological advancements and deployment pace often outstrip regulatory frameworks, leading to potential misuse or inadequate oversight. The integration of these technologies in sensitive sectors such as healthcare, legal services, and emergency management (O'Brien, 2020) requires dynamic governance structures to mitigate risks effectively.

Several potential ethical problems may arise from the deployment and use of AI-powered language technologies. These include privacy violations, involving the risk of unauthorized use or disclosure of sensitive personal data (Weidinger et al., 2021). Systematic biases inherent in training data may result in discriminatory outputs, adversely affecting certain demographic or linguistic groups (Bianchi et al., 2023; Savoldi et al., 2021; Tomalin et al., 2021). Users might misuse or overly rely on automated systems, leading to adverse outcomes, particularly in high-stakes scenarios. Additionally, difficulties in clearly attributing accountability for errors, misinformation, or harm caused by AI-generated outputs present significant ethical concerns (Moniz & Parra Escartín, 2023). Lastly, ethical concerns around data sourcing, including intellectual property rights and consent for data use, underscore the importance of comprehensive ethical frameworks. The ethical dimension also needs to factor in discussion of the carbon footprint involved in the deployment of any technologies served by LLMs (see, for example, Ding and Shi (2024)).

Addressing all the above challenges and proactively mitigating ethical issues are critical to responsibly developing and adopting HCAILT tools and systems, ensuring they serve their intended purpose without compromising communication outcomes, ethical standards or societal trust.

7. Conclusions

This paper has set out the HCAILT model, a domain-specific articulation of Shneiderman's HCAI paradigm that translates the abstract goals of reliability, safety culture and trustworthiness into concrete design levers for multilingual communication systems. By foregrounding two societal drivers (cognitive augmentation and information dissemination), the model positions language technologies not as autonomous tools, but as sociotechnical partners that expand human agency while keeping people firmly in control. We sought to address two research questions within the HCAI framework:

- RQ1. How can reliability, safety and trustworthiness be concretely operationalized in language-technology workflows?
- RQ2. What constitutes a robust framework for the evaluation of reliability, safety, and trustworthiness in language-technology workflows?

For RQ1 we have made explicit suggestions for operationalizing these concepts in different language technology tools by demonstrating how they could be used in two different contexts, healthcare communication and crisis response.

We have also proposed a concrete framework for the evaluation of those concepts within an HCAILT context (RQ2) and we have provided a demonstration system that implements facets of the HCAILT paradigm and demonstrates the feasibility of implementing HCAILT guardrails today, using commodity LLMs and publicly available medical corpora.

The analysis also surfaces limitations that set the agenda for future work. Reliability remains sensitive to data scarcity in low-resource languages; safety culture cannot be engineered without sustained organizational commitment; and calibrated trust depends on user literacy that many public-facing deployments have yet to cultivate. Addressing these gaps will require multidisciplinary collaborations that bring together natural language processing (NLP) researchers, designers, domain experts, regulators and, crucially, end-users. For our use cases analysed, rigorous field trials in hospitals, emergency-operation centers and community hubs are the next empirical step, accompanied by longitudinal studies that trace how HCAILT interventions affect decision quality, equity of access and public trust over time.

Even at this formative stage, the HCAILT model contributes a pragmatic vocabulary and a set of actionable blueprints for scholars, developers and policymakers grappling with the societal consequences of generative AI and AI-powered language technologies. By tying technical guardrails to organizational governance and transparent user interfaces, it charts a path from ethical aspiration to deployable practice.

We invite the community to iterate, critique and empirically test this model, advancing a future in which language technologies are not merely powerful, but demonstrably reliable, safe and worthy of the trust placed in them.

Note

1. Definition by the European Commission: https://digital-strategy.ec.europa.eu/en/policies/language-technologies.

Author contributions

CRediT: **Vicent Briva-Iglesias**: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Sharon O'Brien**: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received.

ORCID

Vicent Briva-Iglesias http://orcid.org/0000-0001-8525-2677
Sharon O'Brien http://orcid.org/0000-0003-4864-5986

References

Ahmad, M. A., Yaramis, I., & Roy, T. D. (2023). Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI.

Antonini, R. (2016). Caught in the middle: Child language brokering as a form of unrecognised language service. *Journal of Multilingual and Multicultural Development*, *37*(7), 710–725. https://doi.org/10.1080/01434632.2015.1127931

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digital Medicine*, *8*(1), 274. https://doi.org/10.1038/s41746-025-01670-7

Baumgarten, S., & Bourgadel, C. (2024). Digitalisation, neo-Taylorism and translation in the 2020s. *Perspectives*, *32*(3), 508–523. https://doi.org/10.1080/0907676x.2023.2285844

Bianchi, F., Fornaciari, T., Hovy, D., & Nozza, D. (2023). Gender and Age Bias in Commercial Machine Translation. In H. Moniz & C. Parra Escartín (Eds.), *Towards responsible machine translation* (Vol. 4, pp. 159–184). Springer International Publishing. https://doi.org/10.1007/978-3-031-14689-3_9

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Bowker, L. (2020). Machine translation literacy instruction for international business students and business English instructors. *Journal of Business & Finance Librarianship*, *25*(1-2), 25–43. https://doi.org/10.1080/08963568.2020.1794739

Bowker, L., & Buitrago-Ciro, J. (2019). *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Publishing Limited.

Briva-Iglesias, V. (2022). English-Catalan neural machine translation: State-of-the-art technology, quality, and productivity. *Tradumàtica: tecnologies de la Traducció*, (20), 149–176. https://doi.org/10.5565/rev/tradumatica.303

Briva-Iglesias, V. (2023). Translation technologies advancements: From inception to the automation age. *La Família Humana: Perspectives Multidisciplinàries de La Investigació En Ciències Humanes i Socials* (pp. 137–152).

Briva-Iglesias, V. (2024). *Fostering human-centered, augmented machine translation: Analysing interactive post-editing* [Doctoral thesis]. Dublin City University.

Briva-Iglesias, V. (2025). Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication [Paper presentation]. In P. Bouillon, J. Gerlach, S. Girletti, L. Volkart, R. Rubino, R. Sennrich, A. C. Farinha, M. Gaido, J. Daems, D. Kenny, H.

Moniz, & S. Szoc (Eds.), Proceedings of Machine Translation Summit XX: Volume 1 (pp. 365–377). European Association for Machine Translation.

Briva-Iglesias, V., & Peñuelas Gil, I. (2025). Simplifying healthcare communication: Evaluating AI-driven plain language editing of informed consent forms [Paper presentation]. In M. I. R. Ginel, P. Cadwell, P. Canavese, S. Hansen-Schirra, M. Kappus, A. Matamala, & W. Noonan (Eds.), *Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL)* (pp. 55–65). Geneva, Switzerland. European Association for Machine Translation.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Sutskever, I. (2020). Language models are few-shot learners.

Cadwell, P., O'Brien, S., Larroyed, A., & Federici, F. M. (2024). Crisis translation maturity model for better multilingual crisis communication. *INContext: Studies in Translation and Interculturalism*, *4*(2), 136–165. https://doi.org/10.54754/incontext.v4i2.98

Chen, Y.-J., Chiou, C.-M., Huang, Y.-W., Tu, P.-W., Lee, Y.-C., & Chien, C.-H. (2018). A comparative study of medical device regulations: US, Europe, Canada, and Taiwan. *Therapeutic Innovation & Regulatory Science*, *52*(1), 62–69. https://doi.org/10.1177/2168479017716712

Conia, S., Lee, D., Li, M., Minhas, U. F., Potdar, S., & Li, Y. (2024). Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs.

Ding, Y., & Shi, T. (2024). *Sustainable LLM serving: Environmental implications, challenges, and opportunities: Invited Paper* [Paper presentation]. 2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC), Austin, TX, USA, 37–38. IEEE. https://doi.org/10.1109/IGSC64514.2024.00016

Dorn, S. (2025). *AI summaries are about to spread across healthcare*. https://www.forbes.com/sites/spencerdorn/2025/02/13/ai-summaries-are-about-tospread-across-healthcare/

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI under the law: The role of explanation. https://doi.org/10.2139/ssrn.3064761

EU (2024). European Union AI Act.

Federici, F. M. and O'Brien, S., editors (2020). *Translation in cascading crises*. Routledge, Taylor and Francis Group.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), e7785. https://doi.org/10.2196/mental.7785

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., & Specia, L. (2020). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 539–555. https://doi.org/10.1162/tacl_a_00330

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Gao, G., Xu, B., Cosley, D., & Fussell, S. R. (2014). *How beliefs about the presence of machine translation impact multilingual collaborations* [Paper presentation]. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, In, CSCW '14 (pp. 1549–1560). Association for Computing Machinery.

Goulet, M.-J., Simard, M., Parra Escartín, C., & O'Brien, S. (2017). La traduction automatique comme outil d'aide à la rédaction scientifique en anglais langue seconde : Résultats d'une étude exploratoire sur la qualité linguistique. *ASp*, (*72*), 5–28. https://doi.org/10.4000/asp.5045

Huang, H., Wu, S., Liang, X., Wang, B., Shi, Y., Wu, P., Yang, M., & Zhao, T. (2023). Towards making the most of LLM for translation quality estimation. In F. Liu, N. Duan, Q. Xu, & Y. Hong (Eds.), *Natural language processing and Chinese computing* (pp. 375–386). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44693-1_30

Kenny, D. (2022). Human and machine translation. In *Machine translation for everyone: Empowering users in the age of artificial intelligence* (Vol. 18, pp. 23). Language Science Press.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.

Koutsouleris, N., Hauser, T. U., Skvortsova, V., & Choudhury, M. D. (2022). From promise to practice: Towards the realisation of AI-informed mental health care. *The Lancet. Digital Health*, *4*(11), e829–e840. https://doi.org/10.1016/S2589-7500(22)00153-4

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, *67*, 653–672. https://doi.org/10.1613/jair.1.11371

Lewis, W. (2010). Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes [Paper presentation]. In F. Yvon & V. Hansen (Eds.), *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation.

Marais, L., Louw, J. A., Badenhorst, J., Calteaux, K., Wilken, I., van Niekerk, N., & Stein, G. (2020). *AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care* [Paper presentation]. 2020 IEEE 23rd International Conference on Information Fusion (FUSION) (pp. 1–8). https://doi.org/10.23919/FUSION45008.2020.9190240

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., Hou, L., Cheng, Y., Liu, Y., Mahdavi, S. S., Prakash, S., Pathak, A., Semturs, C., Patel, S., Webster, D. R., … Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. *Nature*, *642*(8067), 451–457. https://doi.org/10.1038/s41586-025-08869-4

Moniz, H. and Parra Escartín, C., editors (2023). *Towards responsible machine translation: Ethical and legal considerations in machine translation*. Machine Translation: Technologies and Applications (1st ed.). Springer International Publishing.

Montalt, V. (2021). Medical humanities and translation. In *The Routledge handbook of translation and health* (pp. 130–148). Routledge.

Montalt-Resurrecció, V., García-Izquierdo, I., & Muñoz-Miquel, A. (2024). *Patient-centred translation and communication*. Taylor & Francis.

Moorkens, J. (2024). I am not a number: On quantification and algorithmic norms in translation. *Perspectives*, *32*(3), 477–492. https://doi.org/10.1080/0907676X.2023.2278536

O'Brien, S. (2020). Translation technology and disaster management. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (1st ed., pp. 304–318). Routledge.

O'Brien, S. (2024). Human-centered augmented translation: Against antagonistic dualisms. *Perspectives*, *32*(3), 391–406. https://doi.org/10.1080/0907676X.2023.2247423

O'Brien, S. and Federici, F. M., editors (2023). *Translating crises*. Bloomsbury Academic.

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2025). Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling [Paper presentation]. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25 (pp. 1–29). Association for Computing Machinery.

Pava, J., Badi Uz Zaman, H., Meinhardt, C., Friedman, T., Truong, S. T., Zhang, D., Cryst, E., Marivate, V., & Koyejo, S. (2025). Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts | Stanford HAI. Technical report.

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models?

Pym, A., Ayvazyan, N., & Prioleau, J. M. (2022). Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia. *Just. Journal of Language Rights & Minorities, Revista de Drets Lingüístics i Minories*, *1*(1–2), 71–99. https://doi.org/10.7203/Just.1.24880

Rodríguez-Miret, J., Farré-Maduell, E., Lima-López, S., Vigil, L., Briva-Iglesias, V., & Krallinger, M. (2024). Exploring the potential of neural machine translation for cross-language clinical natural language processing (NLP) resource generation through annotation projection. *Information*, *15*(10), 585. https://doi.org/10.3390/info15100585

Rossetti, A., O'Brien, S., & Cadwell, P. (2020). Comprehension and trust in crises: Investigating the impact of machine translation and post-editing [Paper presentation]. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 9–18). European Association for Machine Translation.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. https://doi.org/10.1162/tacl_a_00401

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 1–31. https://doi.org/10.1145/3419764

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, *23*(3), 419–433. https://doi.org/10.1007/s10676-021-09583-1

Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., & Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, *642*, 442–450. https://doi.org/10.1038/s41586-025-08866-7

Valero-Garcés, C. (2025). An approach to languages of lesser diffusion (LLD) and public service interpreting and translation (PSIT) in Spain in the second decade of the 21st century. *FITISPos International Journal*, *12*(1), 201–217. https://doi.org/10.37536/FITISPos-IJ.2025.12.1.407

Vieira, L. N., O'Sullivan, C., Zhang, X., & O'Hagan, M. (2023). Machine translation in society: Insights from UK users. *Language Resources and Evaluation*, *57*(2), 893–914. https://doi.org/10.1007/s10579-022-09589-1

Villarreal, M., MacPherson-Krutsky, C., & Painter, M. A. (2025). Barriers and best practices for inclusive emergency alerts and warnings. *International Journal of Disaster Risk Reduction*, *125*, 105581. https://doi.org/10.1016/j.ijdrr.2025.105581

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from Language Models.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models [Paper presentation]. In *2022 ACM Conference on Fairness Accountability and Transparency* (pp. 214–229). ACM.

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839. https://doi.org/10.1016/j.artint.2022.103839

Xiao, Y., & Yu, S. (2025). Can ChatGPT replace humans in crisis communication? The effects of AI-mediated crisis communication on stakeholder satisfaction and responsibility attribution. *International Journal of Information Management*, *80*, 102835. https://doi.org/10.1016/j.ijinfomgt.2024.102835

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Yeh, P.-L., Kuo, W.-C., Tseng, B.-L., & Sung, Y.-H. (2025). Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. *Current Psychology*, *44*(9), 8133–8145. https://doi.org/10.1007/s12144-025-07359-0

About the authors

Vicent Briva-Iglesias is Assistant Professor in Translation Technology at Dublin City University and adjunct professor in language technologies at McGill University and Universitat Oberta de Catalunya. His main research interests are human-computer interaction and human-centered AI.

Sharon O'Brien is Full Professor of Translation Studies in Dublin City University (DCU), Ireland, and currently Dean of Graduate Studies. Her research centers on the topics of translation technology and human-computer interaction and translation in disaster settings.