**Interpreting SAFE AI Task Force Guidance**

**June 2024**

**AI and Interpreting Services**

**Table of Contents**



## 1. Executive Summary

This comprehensive guide describes fundamental sociolinguistic criteria for the safe, fair and ethical development and implementation of automatic interpreting products using artificial intelligence (AIxAI), also called machine interpreting. A broad cross-section of stakeholders participated in designing this Interpreting SAFE AI framework. This Guidance is for

policymakers, tech companies/vendors, language service agencies/providers, interpreters, interpreting educators, and end-users. This Guidance establishes four fundamental principles as a durable, resilient and sustainable framework for the language industry. The four principles are drawn from ethical, professional practices of high and low resource languages, and are intended to drive legal protections and promote innovations in fairness and equity in design and delivery so that all can benefit from the potential of AI interpreting products.

## 2. Introduction

The SAFE-AI (Stakeholders Advocating for Fair and Ethical AI in Interpreting) Task Force (Interpreting SAFE-AI TF, https://safeaitf.org/), was founded in the summer of 2023 in response to emerging AI products for interpreting based on the use of generative AI and Large Language Models, which we refer to as Automatic Interpreting by Artificial Intelligence (AIxAI).

The Task Force's mission is to establish, disseminate and promote industry-wide guidelines and best practices for accountable design and adoption of AI in interpreting, through facilitating dialogue and action among designers, vendors, buyers, qualified practitioners, end-users, policymakers, and other stakeholders.

### 2.1 The Scope of the Guidance

This Guidance provides a reference point for all stakeholders of the interpreting profession in all working environments (e.g., business, conference, education, legal, medical) in both primary modalities, i.e., spoken and sign languages, with anticipation for modifications that provide meaningful accommodations as needed by diverse individuals.

The Guidance promotes foundational ethical principles and considerations from contemporary professional interpreting theory and practice that could and should be further developed into practical guidelines and recommendations for specific environments, jurisdictions, and modalities.

### 2.2 Background

In December 2023 - January 2024, the Interpreting SAFE-AI Task Force conducted its inaugural research project, a dual-track study consisting of a multi-language perception survey analysis, performed for the task force by CSA Research, and a qualitative study by the independent Advisory Group on AI and Sign Language Interpreting. These efforts fostered a reflective dialogue on AI's implications in interpreting and resulted in the publication of the two comprehensive reports:

- Perceptions on Automated Interpreting (available in English)
- Deaf-Safe AI: A legal Foundation for Ubiquitous Automatic Interpreting (available in English and American Sign Language)

This SAFE AI Guidance is the result of extensive consultations with the stakeholders in the Interpreting SAFE-AI Task Force and the Advisory Group on AI and Sign Language Interpreting. These inaugural guidelines reflect a diverse mosaic of sources and input, including research conducted on behalf of the SAFE-AI TF and emerging regulation and legislation in the European Union, United States and international regulatory bodies, thereby serving as a comprehensive synthesis of multiple perspectives and voices. This Guidance covers the full range of AI-generated, AI-assisted, and Human-supported AI interpreting products and services.

### 2.3 Layers and Degrees of AI Involvement

AI involvement in interpreting varies in the degree of human supervision of the interpreted interaction, from:

1. **Human-generated:** no oversight, to
2. **Human-generated with AI tools:** human generated with AI back-up/assistance: e.g., for terminology specific to the particular context, to ameliorate imperfectly balanced bilingualism, to search for translations of idioms or slang, technical jargon, terms of self-identity; and other spontaneous challenges of co-constructing shared meaning as they arise, to
3. **Human real-time oversight at any point:** supervision with built-in tools enabling the option to intervene/interject for established professional practice purposes of expansion, clarification, correction and other mediations of possible misunderstanding or miscommunication that could result in harm, to
4. **Human review/Quality Assurance:** supervisory review *after* the fact, with immediate mechanisms to redress discovered or suspected errors in meaningful real-time, to
5. **AIxAI (AI-generated interpreting)** – machine interpretation produced by an AI software during real-time human communication without any input of humans during this interaction.

## 3. Ethical Principles for the Use of AI Solutions for Interpreting

### 3.1 Principle 1: End-User Autonomy

**Accountable use of AI technology means that end users are included in the design process from the beginning in order to ensure that AI tools are actually suitable to end user needs. This includes the development pipeline and evaluation as well as interface design, studies of human computer interaction (HCI), and product testing. Once ready for market, AI products that are procured and utilized for interpreting services must include explicit informed consent to accept/decline the use of AI, opt-in/opt out to data collection and storage without penalty, complete transparency, and evidence of adherence to ethical standards**.

Design standards of AI products for interpreting should be created to guide development, governance, and counterfactual* fairness within the system. Organizations deploying AI products must maintain an inventory of the system, collect, process and respond to human feedback, and track corresponding decisions and results. The use of AI tools should be officially recorded in files and records in every instance.

* *Counterfactual Fairness:* captures the intuition that a decision is fair towards an individual if they are the same or belong to a different demographic group (race, gender, age, language group, etc.).

In all settings, some form of each of the following is required:

- Informed Consent to Accept or Decline the use of an AI product, with confidence that a human interpreter will be provided in a timely manner
- Opt In/Opt Out of data collection and storage without penalty
- A mechanism to shift from AI to human interpreting (or vice-versa) at any time
- User friendly grievance process to report errors or harm
- User friendly explanations of "privacy" and "confidentiality" and the layers/degrees of AI involvement in the interpreting process

**3.1.1 Examples:**

- Before a medical consultation – including general healthcare and telehealth – providers and patients are informed about the level of AI involvement in interpreting services that they can choose, highlighting the potential benefits and limitations of each choice, including required quality metrics. They are given the equal choice to accept/decline, ensuring their consent is informed and voluntary and they are comfortable with the mediation of their communication in each and every encounter.
- In legal settings, all parties are briefed on using AI for interpreting services, including how AI interpreters work, the human oversight involved, and the confidentiality measures in place. This process ensures clients understand and give (or withhold) informed consent, and are familiar with their level of autonomous control and rights regarding the service they are receiving. Clients are given the choice to accept and decline, ensuring their consent is informed and voluntary.
- Educational institutions who want to implement AI interpreting tools for, for instance, parent-teacher meetings or IEP discussions, provide a meaningful education process on how these tools assist communication, ensuring parents and guardians make an informed decision about their use. They are given the choice to accept and decline, ensuring their consent is informed and voluntary.

### 3.1.2 Essential Elements for Purchasing and Implementation:

- **Informed Consent Accept/Decline Process:** When proposing to use an AI product, end users should receive comprehensive information on the AI interpreting product(s) and/or service(s), its levels (e.g., fully AI-generated or AI-augmented to support a human interpreter), its capabilities and limitations (via quality metrics), and potential risks or benefits. Highlight how this information will be communicated to users clearly and effectively, including their options if they do not accept. Provide a clear interface and instructions for how users can switch into and out of AI-involved interpreting services at any time. The design should include mechanisms for easily and effectively exercising these options.
- **Opt-in and Opt-out for Data Collection and Storage:** Provide a clear interface and instructions for how users can Opt In or Opt Out of data collection and storage, including the ability to specify purposes (such as 'yes' for language data for low resource languages or 'no' for commercial marketing). Include mechanisms for easily and effectively exercising these options without penalty.
- **Human Oversight:** Identify specific situations where human intervention is crucial in the interpreting process. Discuss how and when human interpreters will be integrated with AI tools to ensure the quality and accuracy of communication outcomes.
- **Compliance with Regulations:** AI in interpreting services must comply with all federal, state, and local regulations, such as HIPAA and FERPA for privacy in the United States and GDPR in Europe. This compliance should be explicitly stated and adhered to in all implementations.
- **Evaluation and Accreditation:** Collaborate with end users to create processes for evaluating and accrediting AI interpreting tools. Outline the criteria and processes for such evaluations to ensure these tools meet the necessary ethical, quality, and regulatory standards.
- **Handling Grievances and Errors:** Establish and publish protocols for addressing any grievances or errors arising from using AI in interpreting services, up to and including definitions of harm and types of harm. This includes corrective processes, quality assurance, and quality control measures, including regular and frequent cycles of continuous improvement.
- **Organizational Workflow Integration:** Offer guidance on integrating AI interpreting services within organizational workflows. This includes decision-making processes for purchasing AI tools, involving stakeholders such as consumer advocacy groups (especially for low resource languages) and interpreting associations, in addition to standard in-company IT, compliance, and risk management.
- **Cultural Competency and Sensitivity:** Highlight the importance of cultural competency in AI interpreting services, acknowledging areas where human interpreters provide value that AI currently cannot, such as cultural nuances and advocacy.
- **Accessibility Considerations:** Guarantee that AI interpreting services are accessible to all individuals, including those with disabilities, by ensuring user interfaces are compatible with assistive technologies, and materials are available in accessible formats, with end-

user capabilities to make visual, text, and audio adjustments in order to achieve equitable communication outcomes.

### 3.2 Principle 2. Evidence of Improving Safety and Wellbeing for End-Users

**As a goal, AIxAI should improve the experience and effectiveness of collaborative communication, not diminish or weaken it. Creation of AI products for interpreting and incorporation of interpreting AI products in human communication must follow or exceed the existing legal and ethical frameworks for provision of interpreting services that are relevant for each particular setting of human communication and jurisdiction.** If an AI product is limited in its ability to meet standards of human interpreting, and a company chooses to deploy it anyway, this limitation must be addressed directly in the Informed Consent process by making all parties aware of the limitations prior to the decision and consent of utilizing it. The option to choose human interpreting must be made available and operational so as not to further inconvenience (or cause harm) to the end user(s).

#### 3.2.1 Examples:

- Title VI of the Civil Rights Act is the foundation for ensuring equal language access to services funded by the federal government for limited-English proficient (LEP) individuals in the U.S.A.
- Section 1557 of the PPACA Rule (45 CFR 92.201) directly address AI and machine translation:

  > § 92.201 Meaningful access for individuals with limited English proficiency.
  > (c) Specific requirements for interpreter and translation services.
  > (3) If a covered entity uses machine translation when the underlying text is critical to the rights, benefits, or meaningful access of an individual with limited English proficiency, when accuracy is essential, or when the source documents or materials contain complex, non-literal or technical language, the translation must be reviewed by a qualified human translator.
- The Americans with Disabilities Act (ADA), the Rehabilitation Act of 1973, and Section 1557 of the Affordable Care Act mandate the provision of ASL interpreters for communication access.
- The Individuals with Disabilities Education Act (IDEA), is a law passed in 1975 guaranteeing free and appropriate public education (FAPE) for children with disabilities. The law specifies types of services and special education offered to meet the needs of a disabled student, such as CART and interpreting services.
- Section 508 of the Rehabilitation Act of 1973, amended in 1998, requires federal agencies to make their electronic and information technology (EIT) accessible to people with disabilities. The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology.

- Principles of equity have international application assuring that disadvantaged individuals receive additional and adequate, compensatory support in order to receive and benefit from the same standard of service available to advantaged individuals.
- Interpreting standards of practice address principles of accuracy, handling omissions and additions, and considerations for cultural context.

- AIxAI standards of quality practice achieve high percentage accuracy, appropriately handle omissions and expansions, and enable consideration of cultural context.

**3.2.2 Essential Elements for Purchasing and Implementation:**

- Identify legal regulations applicable to your setting (areas of practice) and location (jurisdiction) and analyze against public checklists/leaderboards if deployment of the selected AI product would meet these regulations to the same degree that human interpreting does.
- Identify standards of practice for interpreters applicable to your setting (areas of practice) and compare how the selected AI product complies with them. Provide this comparative analysis to end-users as part of the informed consent process.
- Assess possibility and scope of unintended consequences resulting from the deployment of AI products for interpreting and their effect on equity, safety, and dignity for end-users.

**3.3 Principle 3. Transparency of AI Quality for the General Public**

**The principle of transparency refers to:**

- **having explicit policies and procedures that address the implications of developing and using AI for interpreting, including all AI-related aspects and elements: e.g., algorithmic bias, data storage, use and re-use of materials directly and indirectly-related to privacy and confidentiality and contingent upon end-user consent; how to handle requests for data from language service agency/clients, other companies and government agencies; differences in tiers of service between corporate users (e.g., for localization and/or internal business communication such as human resources, mandatory and specialized trainings, daily operations, and leadership meetings) and direct service providers (e.g., government entities, private businesses offering language/interpreting services);**
- **true costs of the translation product are not hidden, and**
- **this information is published (for general SEO discovery) and communicated – along with implications – directly to business/enterprise clients as part of contracting and directly to end-users as part of informed consent.**

The objective* functions of AI for interpreting machine learning should be disclosed to relevant parties – including examples of what types of embeddings are being optimized – and a disclaimer indicated while AI is in use. The disclaimer should explain the key implications of using AI in the given setting.

**Objective Functions* in machine learning/generative AI are the mathematical measure of the quality of the desired optimization. Objective is used in the sense of goal. What is the goal that the AI is being optimized for? Optimization is the process of finding the best, optimal, solution for that goal. The objective function takes data and the specific model's parameters to return a number, i.e., a quantitative score. Objective functions evaluate the parameters in order to adjust their values in order to maximize or minimize achievement of the goal (definition by ChatGPT).

**3.3.1 Levels of Transparency:**

A. *For organizational purchasers of AI solutions*
*B. For end-users of interpreting services*

**3.3.1A Vendors and Purchasers of AI products**

This level of transparency should be granular for both vendors and purchasers.

Vendors:

- Transparent identification of the AI tool as bi-directional or unidirectional; including performance evaluation metrics, general evaluation metrics, model evaluation metrics, confidence and class confusion (current examples listed below in 4.x.x..) for both/all languages.
- Information about dialects, variants, and accents for specified languages and language pairs that have been used for training of AI tools in order to mitigate bias.
- Distinguishing features of the different levels of service that end-users will receive for each layer and degree of AI product AI-generated, AI-assisted, and Human-supported – and different modalities offered: such as automatic caption translation or AI translation of pre-recorded materials.
- Design and provide mandatory trainings for company employees and purchasers.

Purchasers:

- Set policy regarding accountability for fair, safe and ethical levels of service needed for your setting (areas of practice) including definitions of accidental and adversarial harms, and procedures for providing interpreting by humans when customers Decline AI interpreting.
- Establish Quality Assurance procedures and regular reviews, metrics, tools, plans, and schedules for quality improvement, which must be dynamic and continuous.
- Guarantee the ability for end-users to Accept or Decline AI-involved service with human service provision upon request.

- Design and provide mandatory trainings for company employees and users.

**Sample Purchaser's Questions to AI Vendors (Checklist of Disclosures)**

- What languages are available for bidirectional and bimodal interpreting, and at what levels of proficiency?
- What languages are available for unidirectional interpreting?
- Is your AI tool localized to any regional dialects (e.g., Egyptian or Levantine Arabic (among others), Québécois or Swiss French (among others), Brazilian or Mozambican Portuguese (among others), Castilian or Caribbean Spanish (among others), etc.)?
- How much detail will you provide to your clients and to end users about the data (e.g., audio, signed, scraped from the web) used to train your AI tool?
    - Has your AI tool been trained on discourse and terminology pertinent to a specific interpreting setting, e.g., medical, legal, law enforcement, education, business, etc.?
    - Has your AI tool been trained on speech/sign patterns of specific end-users, such as by age, health condition, regional dialect or accent? E.g., if a client wants to use your tool to communicate with children of 6-10 years of age, would you inform the client that your audio recognition was trained on adult voices? Similarly, if the client wants to use your tool for patients from Argentina, will they get information that 80% of training for your tool was done in Mexican Spanish?
- What are the confidence scores for each language and/or regional dialect? How does the AI tool's confidence scoring compare to a human-interpreted session? (If AI interpreting vendor does not offer human interpreting, the purchaser needs to establish their own comparison process.)
- How is feedback incorporated into the machine learning of your AI products? At what points can the tool be adjusted or taught to adjust? Do you offer your clients a mechanism to provide direct feedback to your system? In other words, how trainable by the clients is the system you are selling to them?
- Do you have a mechanism for your AI tool to track "abandonment" by end-users and provide that info to the client? E.g., if an end-user starts a conversation with your AI tool, and then switches to a human interpreter, how will this be tracked and communicated?
- Do you have an embedded evaluation system for end-users ("satisfaction score")?
- How do your end-user "satisfaction scores" for AI interpreting tools compare to human interpreting products? (If an AI interpreting vendor does not offer human interpreting, the purchaser needs to establish their own comparison process.)
- Are your AI products capable of being audited for racial bias? Have they already been audited for racial bias? If yes, what were the results?

### 3.3. 1B For end-users of interpreting services

At a minimum end-users must be informed about:

- the layers and degrees of interpreting service available to choose from: AI-generated, AI-assisted, and Human-supported AI interpreting including their affordances and disaffordances
- their ability and mechanisms to Accept or Decline AI-involved interpreting and instead select human interpreting (or vice-versa)
- the AI's evaluation metrics of bidirectional language translation quality (see 4.4.1),
- Opt-In and Opt-Out for data collection and storage, and
- any necessary training for use.

### 3.4 Principle 4. Accountability by Companies for Errors and Harm to End-users

**The decision to create, sell, and deploy an AI-assisted or AI-generated (AIxAI) interpreting product should be based on sufficient confidence that actual humans with specified responsibilities accept liability for the product's performance, the definition of parameters for implementing the AI product, and consequences of incidents that occur during use. Developers, vendors and purchasers of AI products for interpreting must define and maintain a clear chain of accountability across organizational hierarchies, including any and all managers, engineers, employees, contractors and anyone else participating in the development, deployment and maintenance pipelines.**

Specifically (but not exclusively):

- Liability for risks and harm associated with the use of AI products for interpreting rests with the AI developers, AI vendors and organizations purchasing and deploying AI solutions of any kind. Any divisions of accountability and liability across entities needs to be specified in the contract.
- Purchasers of AI products for interpreting must establish quality assurance policies and procedures that explicitly define limitations of use, and liabilities for misuse, non-disclosure of limitations, or misrepresentation of limitations.
- Prior to deployment, AI interpreting products must undergo validation by qualified human interpreters. This process ensures a high level of accuracy and identifies potential interactional challenges. Validation criteria should be based on the established processes and qualifications of professional human interpreters to ensure AI systems meet or exceed human performance standards.

#### 3.4.1 Examples:

- Potential harms to individuals and groups of individuals from deploying AI products for interpreting may include but are not limited to those resulting from communication

errors, technical failures (such as power outage, internet delivery interruption, poor audio quality), and misuse of AI tools by humans. Definitions of accidental and adversarial harms must be explicit.

- Be cautious of blanket "product use" disclaimers and limitation statements by vendors. E.g., "This product should only be used for simple communication needs." Ask for a clear definition of the "simple communication" concept and create your procedures for how and whom in your organization a context or situation of communication will be designated as "simple."
- Evaluation of AI interpreting product performance requires publication of results from metrics commonly used to evaluate machine learning models, particularly those involved in natural language processing (NLP) such as BLEU, NIST, and METEOR; including Confidence, word error rate (WER), position-independent error rate (PER), recall-oriented understudy for gisting evaluation (ROUGE), class confusion, Top-N accuracy, Jaccard Index, and Validation Loss. Such metrics must, at a minimum, meet quality standards established by ASTM International, NIST, and/or the ISO. The label or claim "State of the Art" is insufficient in and of itself, as this refers to the performance of large language models in experimental conditions, not to the outcomes in real world interactions among living human beings.

### 3.4.2 Essential Elements for Purchasing and Implementation:

- **The risk of accidental and adversarial harm** to humans must be assessed prior to purchasing or deploying any AI interpreting solution, along with their remedies.
- **Laws and regulations** applicable to a particular setting and/or jurisdiction must be applied equally to AIxAI and human interpreting (reference 3.1.2).
- **Human oversight.** Regular scheduled audits by a qualified human interpreter must be conducted and:
    - implemented at multiple points throughout the deployment and utilization of AI interpreting products;
    - address various aspects of AI products for interpreting such as linguistic accuracy, communication effectiveness (based on end-users' reports of satisfaction with desired outcomes), appropriate use by end-users, compliance with internal policies and external regulations;
    - vary in its content so that identified errors are corrected instead of AI simply providing a stored correct answer;
    - confirm that machine learning steadily improves due to identified errors and especially that discriminatory effects and bias are quickly corrected instead of perpetuating systemic harm;
    - assess algorithm management processes to ensure there is no artificial or misleading inflation of quality metrics;
    - verify that end users who Decline are provided with human interpreters;

  - verify that end users who Opt Out of data collection are not recorded;
  - fully document and report appropriately to relevant authorities, clients, and end-users.
- **Metrics** of AI products' performance must be clearly defined and documented in policies and procedures. These metrics should be reported individually per language to avoid concealing deficiencies, especially in the early stages. Language-by-language comparisons are crucial to ensure nuances are accurately captured and not lost in aggregated data. Special attention must be given to differences between high resource languages and low resource languages, with constant effort and evident progress in raising low resource language metrics.
- **Privacy and data integrity.** AI products for interpreting must safeguard data integrity through data controls that protect against unauthorized use or misuse of information/data.
  - Guidelines and policies should include strict limitations on the use of data to prevent any form of misuse, especially including algorithm management. External auditors should confirm the integrity of algorithms on a regular basis.
  - AI products for interpreting must contain mechanisms for end-users who Opt In to comply with applicable regulations (e.g., HIPAA for healthcare participants in the U.S., GDPR in the EU) or any particular end-user wishes to not have their information aggregated or stored within a data warehouse, for any purpose.

**4. Conclusion**

The Interpreting SAFE-AI Task Force invites developers, vendors, language agencies/providers, business organizations, governments and other stakeholders to embrace these principles. By adhering to these standards in a uniform and consistent manner, we collectively cultivate a global society in which AI enriches communication while preserving the autonomy, safety, equitable treatment, and dignity of all involved parties. We encourage you to utilize this Guidance as a compass and fit them to your unique contexts and establish your own customized guidelines and policies.

We urge AI developers and vendors to champion these four principles of

- end user autonomy
- improving safety and wellness for end users
- transparency of quality, and
- accountability for errors and harms

and actively participate in their implementation.

The Interpreting SAFE-AI Task Force will periodically review and refine this Guidance to keep pace with technological advancements, market dynamics, and evolving regulations. Your

feedback, suggestions, and inquiries are invaluable contributions to our ongoing mission to ensure the ethical and responsible integration of AI in interpreting services.

Contact the Interpreting SAFE-AI Task Force at info@safeaitf.org with your comments and suggestions.

**Announcement:**
The Interpreting SAFE-AI Task Force will next begin work on Guidance for Interpreters working with AI interpreting products and support tools.

## 5. Glossary of Key Terms

**Accountability** implies compliance. It involves the clear delineation of what constitutes safe, fair and ethical uses of AIxAI and what constitutes harm. Further, examples of harms should be defined and listed for two categories: accidental and adversarial.

**AI-assisted Interpreting:**

- Explanation: Interpreting services that combine AI tools with human oversight.
- Integration: Ensure AI-assisted ASL interpreting services involve human interpreters to oversee and correct AI-generated interpretations, providing a safeguard against errors.

**AI-assisted Technology:** This term refers to using artificial intelligence (AI) to enhance human tasks. AI helps in tasks like data analysis by quickly processing large amounts of information more efficiently than humans. Although AI does much of the initial work, humans make the final decisions. Examples include smart speakers like Amazon Echo and smart thermostats, which adjust home settings based on your preferences and schedule.

**AI-generated:** This term describes content created entirely by artificial intelligence, with no human involvement during the creation process. For instance, ChatGPT generates responses based on user prompts without human edits during the response creation. It autonomously crafts text that can appear as if a human wrote it, based on its extensive training and algorithms. This process is interactive, guided by the user's input, but the content of each response is directly produced by the AI.

**AI-generated Interpreting:**

- Explanation: Interpreting services fully generated by AI without human intervention.
- Integration: Clearly differentiate between AI-assisted and AI-generated services, providing users with informed choices and ensuring that AI-generated services meet high standards of accuracy and reliability.

**AI Product** is a ready-made software or hardware that uses artificial intelligence to perform specific tasks. It is designed for general use and can be easily purchased and used by many people or businesses. Examples include virtual assistants like Alexa, image recognition apps, and smart speakers.

**AI Solution** is a customized implementation of AI technology designed to solve specific problems for a particular organization. It often involves tailoring AI tools to fit the unique needs and systems of a business. Examples include personalized customer behavior prediction models and AI-powered supply chain optimization.

**AI Solution Parameters**
Information about how an AI software was adjusted for specific use in interpreting. The key parameters for an AI solution for interpreting are (the list below is illustrative, not exhaustive):

- Ownership of software: degree of control a vendor of an AI solution has over how the software has been developed and can be developed in the future (e.g., can new data sets be added, can new features be added, etc.)
- Volume of digital data used for "training" (data analysis) of an AI solution: 200 texts or 2,000,000
- Types of digital data used for "training" (data analysis) of an AI solution: general data or specific to an interpreting setting (e.g., legal, medical, military, K-12, etc.), sources category (e.g., public websites, setting-specific publications, dictionaries), and level human curation of selecting data sets
- Language sources: geographic dialects, including but not limited to a country of origin (e.g., American English or British English) with notation about volumes (e.g., this AI solution has utilized input of 2,000 data sets from Mexico and 20 from Guatemala)
- Audio input and output parameters: characteristics of voices utilized (e.g., age, gender, health/speech impediments, accents within the native speaker language community, accents of non-native speakers of the language)
- Direction/modality of interpreting: bidirectional/bimodal or unidirectional

**AIxAI** is the acronym for Automatic (or Automated) Interpreting by Artificial Intelligence.

**Algorithm management** refers to the systematic process of designing, developing, deploying, monitoring, and maintaining algorithms within an organization or a specific application. It encompasses a wide range of activities aimed at ensuring that algorithms operate efficiently, effectively, and ethically (per ChatGPT).

**Algorithmic Fairness:**

- Explanation: Ensuring AI algorithms do not discriminate and provide equitable outcomes.
- Integration: Implement rigorous fairness checks to ensure AI interpreting tools do not introduce bias against marginalized groups and provide equitable services to all users.

**Artificial Intelligence (AI)** refers to a body of science rooted in mathematics that is itself fundamentally a process of translation into the language of numbers.

**ASTM International** is an international standards organization, formerly known as American Society for Testing and Materials, that develops and publishes voluntary consensus technical standards for a wide range of materials, products, systems, and services, including language teaching, translation, and interpreting.

**Automatic Interpreting** refers to mechanical processes of linguistic translation in place of human interpreting during live human interactions. These are calculated by algorithms based on statistics, and vary in reliability depending on the size and scope of the database for each language. It also includes tools that augment interpreting such as using avatars to mask human

interpreters, ASR captioning with generative predictive text, fingerspelling recognition software, etc.

**Baselines:** Baselines serve as starting points or initial references for comparison in experiments or evaluations. They often represent simple or naive methods that provide a basic level of performance for a given task. Baselines are used to gauge the improvement achieved by more complex or sophisticated algorithms or models. They help researchers or practitioners understand whether proposed solutions are effective or provide significant enhancements over simple approaches.

**Benchmarks:** Benchmarks are established standards or reference points against which the performance of algorithms, models, or systems can be measured. These standards are typically set based on the performance achieved by existing methods or on known results for specific tasks or datasets. Benchmarks provide a point of comparison to assess the effectiveness or efficiency of new approaches.

**Bidirectional Interpreting**
Interpreting that consists of converting meaning in both directions within one human interaction/communication between speakers of two different spoken languages, e.g., from English into Portuguese and from Portuguese into English.

**Bidirectional and Bimodal Interpreting**
Interpreting that consists of converting meaning in both directions within one human interaction/communication between speakers whose languages exist in different articulatory modalities: a sign language and a spoken language, e.g., from American Sign Language (ASL) into English and from English into ASL.

**Bilingual Evaluation Understudy (BLEU)**
BLEU is a popular algorithm for evaluating the quality of machine translation. It is based on the "closeness" between a machine/candidate translation and a reference human translation (Papineni, 2002). Studies have found a high correlation between a high BLEU score and human-judged scores. The BLEU score is a number between 0 and 1, where 1 is considered a "perfect" translation, which is exceedingly rare in practice due to the variability of human translations. The score is calculated by counting n-grams in the reference translation that are reproduced by the candidate translation, with various penalties and bonuses applied. A drawback of BLEU is that it highly depends on tokenization (finding "word" boundaries). Tokenization of ASL can be difficult due to phonemes being used continuously and in multiple contexts. BLEU has seen some use in Signed English systems. If BLEU is reported with a number next to it like BLEU-2, it means that it is counting only 2-grams.

**Briefings** are required in medical, educational, and legal settings prior to use of an AIxAI application. In this briefing, end-users (of both/all languages) are informed of the technical

specs (quantitative quality) of the AIxAI, the levels of AI and/or human involvement available, and a risk/benefit analysis suited to their individual case.

**Clients** are the business, corporation, government entity or other organizational customer of the AI application.

**Collaborative Communication** is a description of the practices (attitudes and behaviors) of the humans involved in interpreted interaction when they are making sincere effort to understand each other both through the interventions of interpreting (by human or machine) and despite the presence of an intermediary who may inject bias (algorithmic or human).

**Confidence**
Loosely, confidence represents the probability that the prediction is correct. Model output can be constrained using confidence cutoffs to prevent unreliable output that the machine is not confident about.

**Confidentiality** refers to knowledge about a person being shared only among those (family members, friends, professional service providers) with an authorized right to know.

**Counterfactual Fairness** ( defined in text, do we want to repeat or expand upon it here? ) captures the intuition that a decision is fair towards an individual if they are the same or belong to a different demographic group (race, gender, age, language group, etc.).

**Cultural Nuances**
Differences in the way individuals in different cultures believe, behave, feel, and express themselves through language, gestures, and body language. These differences impact how messages are perceived and understood. Cultural nuances cannot be inferred from analyzing texts or speech samples.

**Data Privacy:**

- Explanation: Protecting user data used by AI systems in compliance with regulations.
- Integration: Ensure that ASL interpreting tools adhere to stringent data privacy standards, protecting the personal information of users and maintaining confidentiality in sensitive communications.

**Deep Learning:**

- Explanation: A subset of ML that uses neural networks with many layers to model complex patterns.
- Integration: Implement deep learning techniques to enhance the ability of ASL interpreting tools to recognize and interpret subtle gestures and expressions accurately.

**Design justice** ensures that interventions are made first at the product level, focusing on equitable and inclusive design, particularly for marginalized communities, including the Deaf

community and other communities with low resource languages. For instance, in the context of AI for sign language, this means involving Deaf end users at every stage in the development pipeline, from problem selection through data collection, outcome definition, algorithm development and post-deployment considerations.

**Embedding** An embedding is a representation of data (such as words, images, or nodes in a graph) in a continuous vector space. Embeddings capture the semantic meaning or features of the data in a lower-dimensional space, facilitating easier manipulation and analysis (per ChatGPT).

**End-users** are the participants in the interpreted interaction, of both or all languages and modalities (depending how many are in use).

**Equity** recognizes each person has different circumstances and needs, meaning, different groups of people need different resources and opportunities allocated to them in order to thrive. Equity is a measure of value, specifically the value invested in ensuring that individuals receive similar outcomes of fair, safe and ethical treatment regardless of individual traits, characteristics, and/or potentially disabling or discriminatory conditions in a given situation.

**Errors** are mistakes in interpreting that are readily and easily corrected during processes of collaborative communication, including interpreted interaction. The source of errors could be participants from either/any language(s), human interpreters, and/or the automated interpreting algorithms and software. "Errors" imply technical glitches or mistakes that can be corrected, which can minimize the severity and human consequence involved. In contrast, "harms" acknowledge the broader and more serious implications, emphasizing the ethical responsibility to prevent and address the negative outcomes that can affect end-users' well-being, safety, and dignity. Using "harms" highlights the urgency and gravity of these errors, ensuring they are given the necessary attention and resources to be mitigated effectively.

**Ethical** refers to the material behaviors to be promoted, encouraged and maintained through the use of AIxAI.

**Explainability**:

- Explanation: The degree to which the workings of an AI system can be understood by humans.
- Integration: Make machine interpreting AI tools explainable to users, ensuring they understand how decisions are made and the basis for interpretations, thus building trust in the technology.

**FERPA** is the Family Educational Rights and Privacy Act of 1974, a federal law in the United States that protects the privacy of parents and students.

**GDPR** is the General Data Protection Regulation of the European Union (2016). It

- lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
- protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data.

**Generative AI:**

- Explanation**:** AI systems capable of generating natural language responses and interpreting tasks.
- Integration**:** Ensure that generative AI tools for all language pairings for machine interpreting are rigorously tested for accuracy and reliability, and include mechanisms to correct errors in real-time to avoid miscommunication.

**Geographic Dialects**
Varieties of a language that develop in and represent speakers of specific places, regional or local. For example, Mexican Spanish, South American Spanish, Levantine Arabic, Yemeni Arabic, Australian English, Texan American English, etc.

**Harms** are injury or damage to, for example, a person's physical health, emotional or mental well-being, eligibility for programs or services, successful applications or interviews for employment or educational opportunities, and professional reputation within a community of practice. Harms also include group-level patterns of discrimination, stereotyping, and other forms of systemic marginalization.

**HIPAA** is the Health Insurance Portability and Accountability Act of 1996 (United States). It's requirements for the protection of certain health information are specified in the Standards for Privacy of Individually Identifiable Health Information ("Privacy Rule").

**Human-Computer Interaction (HCI)** (Workflow}: The process of evaluating how users interact with the computer interface that allow efficient, effective, and satisfying interactions between humans and computers. It focuses on usability, user experience (UX), accessibility, aesthetics, feedback, consistency, and task efficiency.

**Human-generated:** This term describes content or outputs that are created entirely by humans without the direct involvement of artificial intelligence in the creation process. It involves traditional methods of creation where humans apply their creativity, knowledge, and skills to produce the final product, such as writing a book, painting a picture, or crafting a manual report.

**Interface Design** is the process designers use to create interfaces that are user-friendly, easy to use, accessible, and consistent.

**Interpreting** refers to communication activities during live interaction between two or more human beings who do not fully share a common language.

**Large Language Models (LLMs):**

- Explanation: Advanced AI models trained on extensive datasets to understand and generate human-like text (which may or may not be part of the main products of machine interpreting)
- Integration: Highlight the need for LLMs in pairings of languages for interpreting to be trained on diverse datasets that include dialects and regional variations for spoken and sign language nuances to ensure accurate and contextually appropriate interpretations.

**Limited English Proficient (LEP)** is the federal label for persons whose cognitive fluency is in a language (or languages) other than English. We note that this is an inherently biased label which inherently establishes an unfortunate (discriminatory and stereotyping) linguistic hierarchy.

**Low Resource Languages** "can be understood as less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations." Source: Low-resource Languages: A Review of Past Work and Future Challenges https://arxiv.org/pdf/2006.07264

**Machine Learning (ML):**

- Explanation: AI that learns from data to make decisions and predictions.
- Integration: Use ML to continuously improve pairings of language interpreting tools by learning from user interactions and feedback, thereby increasing accuracy and "fit" to end user's context over time.

**Metric for Evaluation of Translation with Explicit Ordering (METEOR)**
METEOR was developed to address some of the issues with BLEU. It is the harmonic mean of 1-gram precision and recall. Precision is the ratio of true positives to both true and false positives. Recall is the ratio of true positives to both true positives and false negatives (Banerjee, 2005). A study has shown that METEOR may perform better than BLEU or NIST (Lavie, 2009).

**National Institute of Standards and Technology (NIST)**
The NIST metric is based upon the BLEU metric. The main difference is that n-grams are weighted by their informational content. For example, the 2-gram "of the" would have a low weight compared to "broken clock" (Coughlin, 2003).

**Natural Language Processing (NLP):**

- Explanation: AI's capability to understand, interpret, and generate human language.
- Integration: Emphasize the importance of coordinating HCI interface controls with the NLP in low resource language interpreting tools to understand the context and cultural nuances inherent in ASL, ensuring effective communication.

**Objective Functions** in machine learning/generative AI are the mathematical measure of the quality of the desired optimization. Objective is used in the sense of goal. What is the goal that the AI is being optimized for? Optimization is the process of finding the best, optimal, solution for that goal. The objective function takes data and the specific model's parameters to return a number, i.e., a quantitative score. Objective functions evaluate the parameters in order to adjust their values in order to maximize or minimize achievement of the goal (definition by ChatGPT).

**Principles** describe design features in the architecture of AIxAI to enable safe, fair and ethical applications of AIxAI. These features span the algorithmic infrastructure through dashboards and leaderboards to each end-user's specific and particular interface.

**Privacy** refers to the collection of data from a person's online activities associated with the use of AIxAI. This data must always be de-identified for internal and external use and may not be sold to third parties without agreement by and payment to the individuals whose private data is included in any aggregation.

**Title VI of the Civil Rights Act of 1964** prohibits discrimination on the basis of race, color, and national origin in programs and activities receiving federal financial assistance. National origin is understood to include language. Executive Order 12250 required federal agencies to prepare information to affected communities about these protections.

**Speech Recognition Software**
Computer programs that convert human speech to text. This is mostly a step in a complex computer program that then can analyze the produced written text. The main challenge for such programs is to match audio sounds produced by humans with different accents, age-dependent pronunciation (e.g., "child speak"), or speech impediments (e.g., speech patterns of a person who has suffered a stroke) to the correct letter. This challenge is further exacerbated by the quality of audio sound (e.g., how close a person is to the microphone) and presence/absence of additional sounds (e.g., more than one person are speaking, music is played in the background, nearby equipment is producing sounds).

**Transparency:**

- Explanation: Openness about how AI systems function and make decisions.
- Integration: Enhance transparency by providing detailed information about how language pairings using interpreting AI tools are trained, their capabilities, limitations, and the data sources used.

**Unidirectional Interpreting**
Interpreting that consists of converting meaning in one direction from a speaker/signer of one language to listeners/readers/watchers who speak/sign a different language, e.g., from English into Portuguese only. This type of interpreting is deployed when no response from the

listener(s) is expected or needed, e.g., public announcements, lectures to large audiences, news broadcasts, etc.