Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

Balancing Social Impact, Opportunities, and Ethical Constraints of Using AI in the Documentation and Vitalization of Indigenous Languages

Claudio S. Pinhanez , **Paulo Cavalin** , **Marisa Vasconcelos** and **Julio Nogima**
IBM Research, Brazil
{csantosp, pcavalin, marisaav, jnogima}@br.ibm.com

Abstract

In this paper we discuss how AI can contribute to support the documentation and vitalization of Indigenous languages and how that involves a delicate balancing of ensuring social impact, exploring technical opportunities, and dealing with ethical constraints. We start by surveying previous work on using AI and NLP to support critical activities of strengthening Indigenous and endangered languages and discussing key limitations of current technologies. After presenting basic ethical constraints of working with Indigenous languages and communities, we propose that creating and deploying language technology ethically with and for Indigenous communities forces AI researchers and engineers to address some of the main shortcomings and criticisms of current technologies. Those ideas are also explored in the discussion of a real case of development of large language models for Brazilian Indigenous languages.

1 Introduction

Tradition and progress are often in conflict in Indigenous communities and one of its most common battlefields is in strengthening the use of their own languages. We argue in this paper that using *Artificial Intelligence (AI)* and, particularly, *Natural Language Processing (NLP)* technologies to support the documentation and vitalization of Indigenous languages is, possibly, an unusual point of harmony between traditional culture and technological progress. In fact, we see this domain as an area where AI can have a strong and lasting social impact and, at the same time, a domain which enables AI to address some of its key technical challenges and criticisms including the reliance on large amounts of data, the handling of context, a historic foundation on colonial thinking, and the need to address ethical issues.

However, we should first recognize that any discussion about positive impacts of technology must be done with extreme care, as discussed, for instance, by Pal [2017]. We acknowledge that, *"[...] working towards social good cannot be a by-product of a technological intervention, nor can it be a means for us to pat ourselves on the back for another day of a job well done"* [Pal, 2017, pg. 718]. Nevertheless, positive social impacts of AI is one of the central themes of our paper and hereby we assume it to mean using AI technology to contribute to the solution of social and economic problems of under-served and vulnerable communities, according to the needs expressed by them, respecting their social and cultural context, and, whenever possible, in projects led by them.

In fact, creating and deploying technology to be used in Indigenous communities must follow ethical guidelines, as discussed in [Harding *et al.*, 2012; Straits *et al.*, 2012]. Those are in stark contrast with traditional practices of AI, such as reliance on big data, data extractivism, and colonial thinking [Crawford, 2021]. Also, as epitomized by the motto *"Nothing for us without us"* adopted by the Indigenous communities as a premise for any language initiative with them, any work, even in research projects, must be done with the community and for its benefit and in a sustainable manner.

We start this paper recognizing the importance of Indigenous peoples and cultures in the world context. We follow with some data and definitions about endangered languages, a discussion about the value of language diversity, and with an overview of tenants and challenges of documentation and vitalization of Indigenous languages. We then discuss some of the limitations of current NLP technologies to deal with Indigenous languages including issues with *large language models (LLMs)* such as BERT and GPT-3.

After discussing ethical issues and guidelines when working with Indigenous communities, we examine a research initiative conducted by some of the authors of this paper to create language models for 22 Brazilian Indigenous languages and how difficult it was, in practice, to adhere to many of the ethical guidelines while tackling at the same time some of the technical difficulties. This leads to a discussion about trade-offs which researchers and practitioners in this domain are likely to face in order to balance the desire for social impact, the opportunities to advance technical knowledge and tools, and the need to follow ethical guidelines.

We finish by addressing how the domain of documentation and vitalization of Indigenous languages can be of mutual benefit for both AI and the Indigenous communities. For this, we consider some known issues with current AI, e.g. [Crawford, 2021], and explore how creating language-related technology ethically with Indigenous communities may entail the discovery of solutions to long-standing AI problems, such as explainability and common-sense. Following, we dis-

cuss how some problematic AI practices, such as data extractivism and colonial thinking [Birhane and Guest, 2021; Raval, 2019], need to be rethought to allow working with Indigenous communities.

The discussion presented in this paper is closely aligned with three of the *Agenda 2030 of Sustainable Development Goals (SDGs)* of the United Nations[1]: 4 (Quality Education), 10 (Reduced Inequalities), and 16 (Peace, Justice, and Strong Institutions). Also, the *UN Permanent Forum on Indigenous Issues (PFII)* has emphasized the role of *"... protecting indigenous languages and safeguarding traditional knowledge as a crucial element in addressing climate change and other challenges facing the global community today."*[2].

2 Recognizing Essential Peoples and Cultures

As historically has happened in the last centuries, Indigenous peoples continue to be at great risk in the world and particularly in the Americas. During the colonial era many of them were killed, raped, expelled, and prohibited to speak their language. Land grabbing continues to this day in Indigenous territories, leaving entire communities in a permanent state of conflict and at risk of disappearing as peoples.

Indigenous peoples and local communities have also been recognized as fundamental to preserve life diversity on Earth. They are the protectors of 80% of the most bio-diverse areas in the world[3]. Also, Indigenous technological development tends to aid local environmental preservation [Leal *et al.*, 2021]. However, Indigenous peoples are increasingly threatened by legal and illegal bodies operating in or close to their lands, often with violence. In 2019, of the 212 people killed in Latin America for speaking out against environmental damage, 40% were Indigenous people[4].

Not only have the material resources of Indigenous lands been exploited but also their knowledge, for example by pharmaceutical businesses which have capitalize on Indigenous expertise about traditional plant-based medicines. Fortunately, recently those communities have been reclaiming the control of their knowledge and data ecosystems, science, and narratives [Carroll *et al.*, 2020].

Indigenous peoples' wish to preserve and vitalize their cultures and languages is an essential agenda to enable their visibility and a diverse, sustainable, and just world. However, the attitudes of both governments and non-Indigenous societies often make Indigenous people feel their culture and language devalued, pushing them towards the language and culture of the majority groups, often to facilitate transactional communication which incentives their migration to cities.

3 Strengthening Our Linguistic Heritage

There are about 7,000 languages spoken in the world today, of which 4,000 are solely used by the 370 million Indigenous people. Of those, 2,680 are in danger to become extinct by 2100 [Wurm, 2001], part of a historical process of culture loss. For instance, about 85% of the 1,175 Indigenous languages spoken in Brazil in the 16th century have disappeared, together with a reduction from 2.5 million to about 330 thousand people still leaving in traditional ways [Melatti, 2007]. The threat to Indigenous languages has made the UN establish 2022-2032 as the *Decade of Indigenous Languages*[5].

Language endangerment is a continuum and a key determinant of its vitality is whether young children and teenagers speak it or not. Wurm [2001] lists six levels of endangerment: *safe*, *vulnerable*, *definitely endangered*, *severely endangered*, *critically endangered*, and *extinct*, and compiles statistics and geographical maps of endangered languages. Thomason [2015] is a good introductory text about endangered languages, its impact on the individual communities and on humankind. Brenzinger [2008] discusses specific issues of endangered languages across the world.

Indigenous approaches to linguistic data tend to reflect a *"...holistic understanding of language as contextualized language, not simply grammars and dictionaries..."* [Fitzgerald, 2017, p. 291]. They encompass both visible and invisible human and nonhuman presences in the Indigenous environment. Those who study Indigenous collectives have used the term *cosmopolitics* to describe such behavior [Bonfim, 2016], suggesting philosophical, cosmological, and cosmopolitical dialogues between Indigenous and non-Indigenous as a way to break the colonialist inter-ethnic subalternity which is historically rooted in non-Indigenous societies.

Languages are also the most effective record of human linguistic and cognitive evolution [Hale *et al.*, 1992]. Documenting and analyzing languages is as important as archaeology for the understanding of humanity's past. Moreover, languages record distinctive, highly informative ways of thinking and comprehending reality and society [Harrison, 2008]. Maffi [2002] and Loh and Harmon [2014] point out the amount of human knowledge about nature which is intertwined in Indigenous languages and how the extinction of linguistic diversity is also the loss of ancient, important knowledge about biodiversity. Thomason [2015] discusses both the dangers of losing a language for its community and for science. However, as argued by Riverburgh [2013], media often frames endangered languages issues in a way it stimulates complacency and, especially, fatalism.

There are essentially two types of work to avoid the loss of a language: *documentation* and *vitalization*. Documentation is related to processes which collect corpora of utterances, stories, conversations, and written records, both in textual form and in media such as recordings, videos, photographs, etc.; and the creation of grammatical, phonetic, phonological, morphosyntactic, and semantic analysis. Thomason [2015, chapter 6] is a good introduction to how documentation activities have been traditionally done in Linguistics, including important advice on how the researchers should work together with the community and conduct themselves.

Vitalization comprises the activities pursued to maintain

[1] https://www.un.org/sustainabledevelopment/

[2] https://sdg.iisd.org/news/indigenous-peoples-have-a-crucial-role-in-implementing-sdg-16-concludes-permanent-forum/

[3] https://p.dw.com/p/32npT

[4] https://www.globalwitness.org/en/campaigns/environmental-activists/last-line-defence/

[5] https://en.unesco.org/sites/default/files/los_pinos_declaration_170720_en.pdf

and grow the number of speakers of a language and, in particular, the efforts to have children learn it in early age. Notice that vitalization efforts may include work where documentation of the language, gathered in the past or from other sources, is used to help to restore knowledge, enlarge the vocabulary, or recover patterns of speech and accents. When such efforts are done in the context of an extinct or critically endangered language, we use the term *revitalization*. Notable examples are the revitalization of *Hebrew*, from being used almost exclusively in religious ceremonies in early 1900s to being the main language of 7 million people worldwide; of *Catalan* in Spain; and of *Maori* in New Zealand. Pérez et al. [2019] provides a survey of vitalization efforts.

4 AI Technology for Indigenous Languages

The use of computers and AI in both language documentation and vitalization has been quite common. A detailed survey is beyond the scope of the paper but we list here some key works which demonstrate the breadth and originality of this application area of AI and NLP. We also discuss some of the key limitations of current NLP technologies to handle technical aspects of Indigenous languages.

4.1 Related Work

Mager et al. [2018] surveyed work, data resources, and challenges of language technologies for American Indigenous languages. Kuhn et al. [2020] described many different language technology initiatives of the *Indigenous Languages Technology (ILT)* project at the *National Research Council of Canada*, including the construction of corpora for several languages, annotation tools, automatic speech recognition systems, and read-along audiobooks. Neubig et al. [2020] summarized a workshop on the state of use of technology for language documentation in 2019.

In particular, NLP technologies have been used in varied contexts and scenarios of endangered languages. Alavi et al. [2015] discussed whether an automatic conversational system can be used to document languages; Anastasopoulos [2019] explored diverse language tools for language documentation; Anastasopoulos et al. [2020] discussed modern NLP issues with endangered languages; Bird [2018] looked into the specific issue of using mobile technologies; Cruz and Waring [2019] listed linguistic issues of using technology for endangered languages; Everson and Waring [2019] described a platform for community-based description of Indigenous languages; Foley et al. [2018] described the process of building speech recognition systems for language documentation; Katinskaia [2017] presented a language learning system to support endangered languages; Maldonado et al. [2016] described a system for automatic recognition of *Guarani* speech; Martín-Mor [2017] explored the use of technologies for *Sardinian* languages; Maxwell and Bills [2017] discussed how digitizing print dictionaries can help to create data for endangered languages; Mirza [2017] explored social persuasive ubiquitous knowledge systems in the context of the *Maori* language; Simha [2019] explored automatic speech recognition systems; Ubahlet [2021] presented a system to manage corpora of endangered languages; Van Esch et al. [2019] explored future directions to in automatic support for language documentation; Yangarber [2018] explored support for endangered and low-resource languages via e-Learning, translation, and crowd-sourcing; and Zuckermann et al. [2021] studied a web platform for revival and documentation based on community engagement. Finally, the ComputEL annual workshop is a good resource of real, field applications of technology to endangered languages[6].

4.2 Limitations of Current NLP Technologies

Given the outstanding advances of modern NLP for high-resource languages such as English, it is crucial to understand if such advances hold for Indigenous languages given that current language models cover at most 200 of the 7,000 languages in the world, of which none are Indigenous languages [Costa-jussà *et al.*, 2022]. Since the remaining set comprises many ultra low-resource languages, the lack of digital documentation (or any documentation at all) makes it very difficult to conduct test cases to understand better the applicability of such language models for low-resource languages as discussed by Hederich et al. [2020].

As many Indigenous languages are also endangered languages [Anastasopoulos *et al.*, 2020], a lot of additional challenges are expected, such as the lack of written documents and literacy, reliance on older speakers, and infrequent use of language [Rangel, 2019]. With the only exception of podcasts which have been hosted by Indigenous and spoken in their languages, resources such as PDF files, books, and TV shows which, in the worst-case-scenario could be digitized and transcribed, are scarce [Hämäläinen, 2021].

Consequently, one direction to be followed is *zero-shot learning* approaches such as the one applied on 10 Indigenous North-American languages by Ebrahimi et al. [2022] or the data augmentation method used for speech recognition by Simha [2019]. In addition, pre-processing techniques for low-resource languages face additional challenges [Wisniewski *et al.*, 2020] and it is reasonable to think that before building Indigenous language models we may need to take a step back and investigate more elementary techniques.

Another point to be considered is the use of *large language models (LLMs)* such as BERT, GPT-3, and T5, one of the main trends of modern NLP which has been making considerable impact in applications such as text classification, comprehension, generation, and translation in recent years [Han *et al.*, 2021]. LLMs are trained on a plethora of textual data and, despite the main focus on the English language, we have seen their expansion to other languages [Xue *et al.*, 2021; Costa-jussà *et al.*, 2022].

Considering that it might be very difficult to collect enough data to train the billions of parameters of an LLM for most languages, recent works explore the adaptation of such models from one language to another. Muller et al. [2021] used pre-trained language models for new, unseen languages. The best results when adapting BERT to those languages were achieved by translating unseen languages to a language with similar structure but with more resources. Alnajjar [2021] explored an analogous idea to create word embeddings where

[6]https://computel-workshop.org/

the words were individually translated to a language with more resources. Both works relied on previous evidence that language models can generalize to languages with similar structure [Hu *et al.*, 2020]. Such research has indicated that it is possible to make use of modern NLP in very low-resource languages when at least some data is available.

5 Working with Indigenous Communities

We now focus on ethical issues when performing research with Indigenous communities and languages, what is subject of specific guidelines and legal issues. Straits et al. [2012] presented a set of guidelines on how to engage in research with Native US American communities based on 11 principles: native-centrism, respect, self-reflection and cultural humility, authentic relationships, honoring of community time frames, building on strengths, co-learning and ownership, continual dialogue, transparency and accountability, integrity, and community relevance. Those guidelines are applicable both to traditional research and cases where technology development and deployment is involved.

Besides the ethical considerations, there are specific legal and regulatory procedures which have to be followed in different countries and when working with specific Indigenous communities. Harding et al. [2012] discussed issues related to data sovereignty, consent, and intellectual property (IP) rights related to tribal research in the US and alerted for many specific characteristics of agreements between research institutions and Indigenous communities. For instance, there must be special procedures for informed consent processes and involvement of community members in defining exposure and risk to the community. Additionally, specific provisions are needed related to data ownership and sovereign rights since those concepts may be understood quite differently by the community from what they often mean in a Western culture. Harding et al. [2012] also provided a comprehensive list of codes of ethics and IP rights adopted by communities in the US and Canada. Similarly, Sahota [2007] discussed the need of research regulation in US American Indian and Alaska Native communities and challenges to establish this regulation.

5.1 Engagement Guidelines

In the specific context of the UNESCO Decade of Indigenous Languages, including the use of technology for documentation and vitalization, there is a proposed engagement framework described by the *Los Pinos Declaration*[7]. The central tenant is inspired in a slogan created by disability activists, *"Nothing for us without us,"*. Also, the declaration proposes a set of specific guidelines for work to be conducted in the program including the provision of *"...access to sustainable, accessible, workable and affordable Indigenous knowledge records, language technologies and media."*

The Los Pinos Declaration states that *"Digital technologies [...] should contribute to the intergenerational transmission, preservation, revitalization, creation and promotion of Indigenous languages ..."* but warns that *"... Indigenous languages will require substantial involvement of Indigenous peoples, particularly Indigenous women, youth and elders, through their own representatives and institutions."*

A more specific guide for AI-related work, including tool and technology design methodologies, was proposed by the *The Indigenous Protocol and Artificial Intelligence (A.I.) Working Group* [Lewis *et al.*, 2020] as a result of two workshops with Indigenous leaderships, linguistic professionals, and computer researchers. Nevertheless, although large AI conferences have had workshops dedicated to Indigenous contexts, such as in *NeurIPS'20* and *'21*, *ICML'21*, and *ACL'22*, the discussion of ethical guidelines when working with Indigenous peoples is still limited in the AI community.

5.2 Listening to the Communities

Academic collaborations with and between Indigenous people and its associated difficulties have been explored at length. According to Linda Tuhiwai Smith, *"... [the word research] is probably one of the dirtiest words in the Indigenous world's vocabulary."* [Smith, 1999, p. 1]. To address those issues, she proposed that *relational accountability* could be used since is inherent to Indigenous ways of doing. *Relational accountability* [Wilson, 2008] states that relationships are important in research and that all the parties are responsible for them. It is based on Indigenous ways of life and emphasizes the importance of relationships over reliability and validity. It dismisses any desire for impartiality as unrealistic and instead emphasizes accountability for the relationships in which one conducts research.

To address the complex context of academic collaborations between Indigenous and non-Indigenous persons, it is fundamental to involve Indigenous members since the beginning of a research or technological project. However, funds often just come with the acceptance of a clear proposal. How to involve an Indigenous without being able to promise a minimum return? How far advanced should a project be before involving Indigenous people? Would it not be better if the Indigenous had the power to pick the researchers they want to work with?

In fact, the authors of this work believe that Indigenous peoples want to be heard for what they have to say, not for what we want to hear or document. What is the priority of a documentation and vitalization project in an a particular Indigenous community? Only being humble and with a very close understanding of the systemic oppression, traumas, and wounds which any Indigenous community has experienced we, as researchers, designers, and developers, might find ways to co-create projects which benefit all communities: Indigenous, Academics, and Academic Indigenous.

6 Balancing Social Impact, Technology Needs, and Ethics Constraints in Practice

As discussed, achieving positive social impact through the use of advanced technical tools must follow ethical guidelines and practices. However, the practice does not to provide such clear-cut cases of win-win situations, pushing developers and communities to balance pros and cons in the process of development and deployment of specific systems and solutions.

We explore here this "balancing act" of creating and using technology for Indigenous languages in the context of the

[7] https://www.worldindigenousforum.com/products/los-pinos-declaration-chapoltepek-outcome-document

development of LLMs for *Brazilian Indigenous languages (BILs)* performed by some of the authors of this paper. The main goal of that work was to develop LLMs which could act as main NLP engines for most BILs and possibly be used as a platform for a variety of tools. The hope was to overcome the need of large amounts of training data for each BIL which is not feasible given that most of them have less than 100 speakers [IBGE, 2010]. Nevertheless, many of those languages have strong similarities in vocabulary and grammatical structure making the LLM approach very attractive, as shown in some previous works [Ebrahimi *et al.*, 2022; Adelani *et al.*, 2022; Pine *et al.*, 2022].

To test this idea, we developed a prototype LLM for BILs based in the trained *mBERT* model [Devlin *et al.*, 2019], which is an encoder-only LLM trained with masked sentences for self-supervised language modeling. *mBERT* was preferred over *BERT* [Devlin *et al.*, 2019] since mBERT is pre-trained with Portuguese, Spanish and other 102 languages what may favor the performance of similar low-resource languages in downstream tasks [Zoph *et al.*, 2016].

The training of the system was based on translations of *The Bible*, publicly available from websites, to 51 BILs and 3 Western, high-resource languages, *Portuguese*, *Spanish*, and *English*. These additional languages were used to better understand their (dis)similarities compared with the BILs. The resulting LLM was tested in a language identification task and achieved accuracies of around 95%, validating the hypothesis that LLMs can be viable foundational tools.

The process of developing this first prototype has arguably violated many of the ethics guidelines discussed before. First, we did not involve any Indigenous community neither in the decision to create the prototype nor in the technical choices of the project. Second, inclusion of each language did not involve the permission of the corresponding peoples and communities. Third, we used, as our data source, a text from a religion which not only is alien to the Indigenous cultures of Brazil but has also often been associated with colonization efforts and de-culturation practices. Fourth, it is unclear whether the NLP tools which can be created with the LLM are needed or wanted. Fifth, and perhaps most importantly, we have not considered beforehand the governance model of the LLM nor discussed it with Indigenous leaderships.

Would have made sense to have it done otherwise? As AI and NLP experts, we know how dependent current technology is from high volumes of training data. Unless there is a viable solution for the low data problem, it is hard to imagine how basic tools such as translators, audio and video transcriptors, and writing assistants can be created. We could see the possibility of actual social impact if such tools were available but before engaging communities and use their time, effort, and expertise, we felt we needed some evidence that there was something concrete to offer. In such conditions, sidelining the guidelines may make sense provided that *damage containment procedures* are put in place, as discussed next, to avoid undesirable consequences. That is the essence of the balancing act we believe has to be done in this area.

The first, most essential part of the containment process is for everyone involved in the task to be fully aware of the potential "monster" being created. The team was reminded constantly of the negative aspects of training a LLM for Indigenous languages with biblical sentences and that the model should never be used in actual tools. Second, as part of the containment process, there are no plans to release publicly such a model, nor should be, not even among researchers. It is part of an experiment used to test the feasibility of the approach. Third, it should not be used in prototypes of tools, especially in any actual system deployed or tested in a community, since it may generate inappropriate or offensive language, unless with express authorization from the community. Fourth, its existence and shortcomings should be disclosed to the Indigenous communities and other stakeholders.

However, this is just the first step of the balancing act. We are now more confident, based on the results of the "monster" LLM prototype for BILs, that we can actually develop AI and NLP tools for languages with a small number of speakers. At the time of the writing of this paper, we have actually started a collaboration with an Indigenous community close to our laboratory, focused on the creation of writing-support tools for high-school native speakers. With them, we are in a co-design and development process leading to a tool which they and their teachers believe is appropriate and needed. For that, we are also engaging Indigenous languages experts and Academics. In parallel, we are also contacting the work group in Brazil which leads the activities related to the UNESCO Decade of Indigenous Languages, aiming to develop a work and governance model to enable the development of LLMs for BILs which is sustainable both technically and ethically.

Each phase of both processes is likely to require a new set of ethical trade-offs and damage containment procedures. The key is to avoid that the combination of impact, technical, and ethical constraints paralyzes the experimentation, learning, innovation, testing, and deployment processes. It is complex balancing act to allow progress to happen.

7 Indigenous Languages as a Domain for AI

Supporting efforts for documentation and vitalization of Indigenous languages is a great example of using AI technology to contribute to the solution of language-related problems of Indigenous communities. Of course, this can only be accomplished by following the needs expressed by them and respecting their social and cultural context and by following ethical guidelines, often in difficult balancing acts as discussed before. But we believe that the creation, development, and deployment of NLP in the context of Indigenous languages is also a great domain to explore some of the key challenges facing AI today; and a domain where AI will only succeed if it addresses some of its questionable practices.

7.1 Forcing AI to Face Key Technical Challenges

AI technology faces many structural technical challenges, including addiction to data, over-confidence on LLMs, limited explainability, handling of context, and common-sense reasoning. We next argue that Indigenous languages are likely to drive forward technology addressing those challenges.

Vanquishing the Addiction to Data

Most of today's AI and NLP algorithms are based on massive amounts of data, including words and sentences for input

embeddings, grammatical and morphological corpora, transcribed speech, and paired sentences for translation. This has created the notion of *high-resource languages* (such as English, Chinese) and *low-resource languages* (Italian, Portuguese). The amount of AI-suitable data available for most Indigenous languages is significantly smaller than of the low-resource languages. In the case of endangered Indigenous languages, the problem is compounded by lack of written documents and a limited number of speakers [Rangel, 2019].

The NLP community should see those limitations as an opportunity to develop methods, algorithms, and data usage patterns, such as *zero-shot-learning* [Ebrahimi *et al.*, 2022], for such contexts. Also, some recent techniques have elicited data from multiple documents [Maxwell and Bills, 2017; Yangarber, 2018] or augmented data [Kruthika, 2019] and used those in similar contexts such as professional languages or in minority dialects of high-resource languages.

Rethinking Large Language Models

Considering the data limitations of Indigenous languages as just discussed, building LLMs for them, as we do today for English and Chinese, will be difficult. At this point, it is not clear whether multi-lingual models trained with multiple languages [Xue *et al.*, 2021], based on high availability of web data of low-resource languages, are able to perform well with Indigenous languages. However, it seems unlikely, given those languages tend to have distinct linguistic roots from the most common languages in the web.

However, this can be an opportunity to rethink LLMs to make them incorporate the historic and temporal perspectives of human language contained in Indigenous languages. A possible approach is, for instance, to train LLMs to use the rich heritage structure of languages to derive key relationships from languages of the same linguistic branch.

Addressing the Limits of Explainability

The black box nature of most ML algorithms in use today is, in spite of all the efforts of the *Explainability* agenda [Adadi and Berrada, 2018], still an obstacle to any application of AI which needs to explain itself. Arguably, non-explainable systems violate at least two of the ethical guidelines discussed before [Straits *et al.*, 2012]: self-reflection and cultural humility; and transparency and accountability. Moreover, limited understanding of the NLP systems may hamper adoption and unexplainable mistakes are likely to trigger distrust.

A possible way to avoid such consequences is to assure that any AI-based system is built with and, whenever possible, by the Indigenous community. More use of co-design and development is a change in AI practice which is increasingly solicited in many contexts and fundamental, in our view, when dealing with Indigenous groups. However, co-design and development may require support technology in the form of simple and easy tools, more transparent algorithms, and a non-Indigenous technical workforce trained to be respectful of Indigenous knowledge, culture, and needs.

Operating in Non-Traditional Contexts

Indigenous communities often live in non-urban areas, in special relationships with nature, animals, and plants. Creating AI technology which is able to function well in such contexts is an enormous challenge. We see this an opportunity where AI technology is forced to address one of its long-standing problems that is the assumption of operating in a *closed world*. This assumption is often used, albeit hidden, in visual and language classification tasks, since certain patterns of behavior are expected in the unknown situations. However, those patterns tend, in fact, to assume that the context is urban and of a Western culture. The use of NLP in Indigenous contexts will almost often break those assumptions, triggering a need of AI systems which are able to represent context more explicitly and do recognition and reasoning in an *open world*, that is, with algorithms which assume that there are entities unseen before and concepts unknown to them.

Reasoning with Different Common-Senses

Common-sense reasoning is one of the oldest and hardest problems in AI and a key difficulty is its embedding in the knowledge being used by AI systems. Indigenous cultures add more complexity since they may have different ways to explain the happenings and facts of the world. For example, *naive physics* is intertwined with pseudo-scientific Western views of the world and Indigenous cultures often rely on different premises to explain how things behave. This may lead AI and NLP researchers to address common-sense head-front and to find ways to represent it in culturally-dependent forms. In AI, common-sense is often portrayed as universal, such as in the *Never-Ending Language Learner (NELL)* [Mitchell *et al.*, 2018], although those views contradict linguistic and anthropological evidence [Barnhardt and Kawagley, 2005].

Reducing Power and Connectivity Requirements

There is also another contextual dimension which is connected to the infrastructure challenges of non-urban environments and, in many cases, of forests and desolated areas. Applications in such contexts can not rely in constant, if at all, access to the Internet and even on reliable electric power. Additional constraints on algorithms, in terms of connectivity and power consumption, are likely to be needed. Such considerations can benefit AI in general, given the need to address planet sustainability and climate change.

7.2 Overcoming Questionable Practices of AI

Documentation and vitalization of Indigenous languages is also a domain where we do not believe AI is going to succeed if some of its ethically doubtful and unsustainable practices [Crawford, 2021] continue to be used, such as reliance on data extractivism, colonialist views, and the development of unsustainable, unfair, and unjust systems. We now explore how Indigenous languages may entice AI to overcome some of its problematic practices.

Abandoning Data Extractivism

Modern AI has flourished based on the use of vast collections of data, most of it collected without the consent of or the benefit of their owners or creators. The behavior of some researchers and many enterprises have been similar to colonial *extractivist* practices, where resources were gathered by occupiers without recompense, often violently. Moreover, the

"... myth of data collection as a benevolent practice in computer science has obscured its operations of power, protecting those who profit most while avoiding responsibility for its consequences." [Crawford, 2021, pg. 121].

This culture of appropriation of data is, as discussed before, inadequate and unethical in the Indigenous context and, in some countries, outright illegal [Harding *et al.*, 2012]. Linguistic data from Indigenous peoples requires individual and community consent and a complex discussion with the communities about ownership of data, copyrights, and fair compensation for their language and culture. Also, mechanisms for maintaining provenance, controlling access, and providing compensation need to be developed and incorporated into the data and algorithms. Such methods, fundamental in the case of Indigenous languages, are likely to be also welcomed by other communities as important progress in AI.

Decolonizing AI

"Computational and cognitive sciences [...] are built on a foundation of racism, sexism, colonialism, Anglo- and Euro-centrism, white supremacy, and all intersections thereof ..." [Birhane and Guest, 2021]. We should be careful to not allow the nice sound of *"AI for Good"* to become a new, technologically updated version of the Kipling's infamous *"White Man's burden"* of 1899. Besides, given the misgivings of Indigenous communities, for instance, in the handling of Indigenous people's DNA [Caron *et al.*, 2020], there is a natural distrust about technologies which are not rooted in their own practices and over which they can easily lose control.

We believe the use of AI in documentation and vitalization of Indigenous languages should follow *decolonizing* or *decolonial* practices [Alvarado *et al.*, 2021; Raval, 2019; Ali, 2013] including the Indigenous leadership of the process, inclusive and diverse co-designing and co-working, and the centrality of the community culture in terms of needs, process, and delivery. The aim should be to bring into dialogue Western and Indigenous knowledge in order to develop socially-just approaches to the application of AI technologies. Those issues framed the workshop described by Lewis et al. [2020], a multi-cultural workshop to ethically create tools to support Indigenous languages.

Creating Socially Just and Sustainable Technologies

AI, NLP, and Data Science have a track record of applications which are not transparent to their users and with social biases [Eubanks, 2018]. Often, AI is used in contexts to *"...further skew power imbalances by placing more control in employer's hands."* [Crawford, 2021, pg. 219]. We see the domain of Indigenous languages as an unique opportunity for the AI community to change. Not only traditional methods are likely to raise barriers of distrust, but they will also collide head-on with Indigenous views about the expected relationships among people and between people and things.

Moreover, for AI to succeed in this scenario, Indigenous, Academics and Technologists have to practice their communication to understand the project to be built, to acknowledge the value and contribution of each one in the project, and to see value other than financial or academic. Systems will only be effective and sustainable when led, built, and run with Indigenous people, and usable even with limited technical support. Notice this is the condition where most people believe computer systems should be deployed in any community.

8 Final Discussion

Indigenous languages are precious knowledge resources and one of the main sources of historical understanding of the development of human cognition. They are also a key part of the identity and culture of Indigenous peoples who have been the foremost victims of the Western civilization. We believe AI technology can have an important, positive impact supporting those peoples to document and vitalize their languages.

However, as explored in the development of the LLM for Brazilian Indigenous languages, balancing social impact, technological opportunities, and ethical constraints is quite difficult in practice. We see here not only the need of more extended discussion about admissible trade-offs and practices but the need of elaboration of design and development guidelines for the use of language technology in, with, and by Indigenous communities, created and mutually agreed by Academics, Technologists, Linguists, and Indigenous peoples.

Moreover, we believe that working with Indigenous languages may benefit AI in many forms. It may push the technological boundaries of AI to overcome its addiction to data, rethink large language models, address the limits of explainability, operate in non-traditional contexts, use a culturally-informed view of what common-sense is, and address climate change. At the same time, AI has to abandon some of its questionable practices, such as data extractivism, colonization mentality, and lack of fairness and community sustainability. Documentation and vitalization of Indigenous languages has this unique quality of pushing AI to be better in terms of technology and ethics at the same time.

Acknowledgments

This work is part of a collaboration with professors and students of the University of São Paulo, in particular with Luciana Storto, Thomas Finbow, Alexander Cobbinah, and Sarajane Peres. Earlier versions benefited from ideas by Debora Leal and Alexander Rademaker and later versions from the work of Pedro Domingues and Priscila Mizukami. We are also in debt to people and leaders from Indigenous communities in Brazil with whom we have been interacting during the last year. This work results from the collaboration of IBM Research with the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation.

References

[Adadi and Berrada, 2018] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[Adelani *et al.*, 2022] D. Adelani, J. Alabi, A. Fan, J. Kreutzer, et al. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proc. of ACL'22*, July 2022.

[Al-Najjar, 2021] K. Al-Najjar. When word embeddings become endangered. *CoRR*, abs/2103.13275, 2021.

[Alavi *et al.*, 2015] S. Alavi, J. Brixey, and D. Traum. Can we use a spoken dialogue system to document endangered languages? In *Dialog for Good Conference*, 2015.

[Ali, 2013] M. Ali. Towards a decolonial computing. In *Proc. of CEPE*, pages 28–35. International Society of Ethics and Information Technology, 2013.

[Alvarado *et al.*, 2021] A. Alvarado, J. Maestre, M. Barcham, M. Iriarte, et al. Decolonial pathways: Our manifesto for a decolonizing agenda in HCI research and design. In *Extended Abstracts of CHI'21*, 2021.

[Anastasopoulos *et al.*, 2020] A. Anastasopoulos, C. Cox, G. Neubig, and H. Cruz. Endangered languages meet modern NLP. In *Proc. of COLING'20*, 2020.

[Anastasopoulos, 2019] A. Anastasopoulos. *Computational tools for endangered language documentation*. PhD thesis, University of Notre Dame, 2019.

[Barnhardt and Kawagley, 2005] R. Barnhardt and A. Kawagley. Indigenous knowledge systems and Alaska Native ways of knowing. *Anthropology & education quarterly*, 36(1):8–23, 2005.

[Bird, 2018] S. Bird. Designing mobile applications for endangered languages. In *Oxford Handbook of Endangered Languages*, pages 842–861. Oxford University Press, 2018.

[Birhane and Guest, 2021] A. Birhane and O. Guest. Towards decolonising computational sciences. *Women, Gender and Resesarch*, 2021:60–73, 2021.

[Bonfim, 2016] E. Bonfim. Gramáticas cosmopolíticas: o caso Bakairi. In *Proc. do Seminário Ibero-americano de Diversidade Linguística*, 2016.

[Brenzinger, 2008] M. Brenzinger, editor. *Language Diversity Endangered*. De Gruyter Mouton, 2008.

[Caron *et al.*, 2020] N. Caron, M. Chongo, M. Hudson, L. Arbour, et al. Indigenous genomic databases: pragmatic considerations and cultural contexts. *Frontiers in Public Health*, 8:111, 2020.

[Carroll *et al.*, 2020] S. Carroll, I. Garba, O. Figueroa-Rodríguez, J. Holbrook, et al. The CARE principles for Indigenous data governance. *Data Science Journal*, 19(1):43, 2020. Number: 1.

[Costa-jussà *et al.*, 2022] M. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[Crawford, 2021] K. Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

[Cruz and Waring, 2019] H. Cruz and J. Waring. Deploying technology to save endangered languages. *ArXiv*, abs/1908.08971, 2019.

[Devlin *et al.*, 2019] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of ACL'19*, June 2019.

[Ebrahimi *et al.*, 2022] A. Ebrahimi, M. Mager, A. Oncevay, et al. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proc. of ACL'22*, 2022.

[Eubanks, 2018] V. Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[Everson *et al.*, 2019] R. Everson, W. Honoré, and S. Grimm. An online platform for community-based language description and documentation. In *Proc. of ComputEL'19*, 2019.

[Fitzgerald, 2017] C. Fitzgerald. Understanding language vitality and reclamation as resilience: A framework for language endangerment and 'loss'. *Language*, 93(4):e280–e297, 2017.

[Foley *et al.*, 2018] B. Foley, J. Arnold, R. Coto-Solano, G. Durantin, et al. Building speech recognition systems for language documentation: the CoEDL endangered language pipeline and inference system (ELPIS). In *Proc of SLTU'18*, 2018.

[Hale *et al.*, 1992] K. Hale, M. Krauss, L. Watahomigie, A. Yamamoto, C. Craig, M. Jeanne, and N. England. Endangered languages. *Language*, 68(1):1–42, 1992.

[Hämäläinen, 2021] M. Hämäläinen. Endangered languages are not low-resourced! In M. Hämäläinen, N. Partanen, and K. Alnajjar, editors, *Multilingual Facilitation*. University of Helsinki, Helsinki, Finland, 2021.

[Han *et al.*, 2021] X. Han, Z. Zhang, N. Ding, et al. Pre-trained models: Past, present and future. *AI Open*, 2021.

[Harding *et al.*, 2012] A. Harding, B. Harper, D. Stone, C. O'Neill, et al. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. *Environmental health perspectives*, 120(1):6–10, 2012.

[Harrison, 2008] K. Harrison. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press, 2008.

[Hedderich *et al.*, 2020] M. A Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

[Hu *et al.*, 2020] J. Hu, S. Ruder, A. Siddhant, G. Neubig, et al. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020.

[IBGE, 2010] IBGE. Características gerais dos indígenas: Resultados do universo. *Censo Demográfico 2010*, 2010.

[Katinskaia *et al.*, 2017] A. Katinskaia, J. Nouri, and R. Yangarber. Revita: a system for language learning and supporting endangered languages. In *Proc. of the workshop on NLP for CALL and for Language Acquisition*, 2017.

[Kruthika, 2019] P. Kruthika. Improving automatic speech recognition on endangered languages. Master's thesis, Rochester Institute of Technology, 2019.

[Kuhn *et al.*, 2020] R. Kuhn, F. Davis, A. Désilets, E. Joanis, et al. The Indigenous languages technology project

at NRC Canada: An empowerment-oriented approach to developing language software. In *Proc. of COLING'20*, pages 5866–5878, 2020.

[Leal *et al.*, 2021] D. Leal, A. M Bustamante, M. Krüger, and A. Strohmayer. Into the mine: Wicked reflections on decolonial thinking and technologies. In *C&T'21*, 2021.

[Lewis *et al.*, 2020] J. Lewis, A. Abdilla, N. Arista, K. Baker, et al. *Indigenous protocol and artificial intelligence position paper*. Indigenous Protocol and Artificial Intelligence Working Group, 2020.

[Loh and Harmon, 2014] J. Loh and D. Harmon. *Biocultural diversity: threatened species, endangered languages*. WWF Netherlands, 2014.

[Maffi, 2002] L. Maffi. Endangered languages, endangered knowledge. *International Social Science Journal*, 54(173):385–393, 2002.

[Mager *et al.*, 2018] M. Mager, X. Gutierrez-Vasques, G. Sierra, and I. Meza. Challenges of language technologies for the Indigenous languages of the Americas. In *Proc. of COLING*, page 55—69, 2018.

[Maldonado *et al.*, 2016] D. Maldonado, R. Villalba Barrientos, and D. Pinto-Roa. Eñe'e: Sistema de reconocimiento automático del habla en Guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016).*, 2016.

[Martín-Mor, 2017] A. Martín-Mor. Technologies for endangered languages: The languages of Sardinia as a case in point. *MTm*, 9:365–86, 2017.

[Maxwell and Bills, 2017] M. Maxwell and A. Bills. Endangered data for endangered languages: Digitizing print dictionaries. In *Proc. of ComputEL'17*, 2017.

[Melatti, 2007] J. Melatti. *Índios do Brasil*. Edusp, 2007.

[Mirza, 2017] A. Mirza. *Design and implementation of social persuasive ubiquitous knowledge systems to revitalise endangered languages*. PhD thesis, Auckland, 2017.

[Mitchell *et al.*, 2018] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, et al. Never-ending learning. *Communications of the ACM*, 61(5):103–115, 2018.

[Muller *et al.*, 2021] B. Muller, A. Anastasopoulos, and otheres. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proc. of ACL'21*, 2021.

[Neubig *et al.*, 2020] G. Neubig, S. Rijhwani, A. Palmer, J. MacKenzie, et al. A summary of the first workshop on language technology for language documentation and revitalization. *arXiv preprint arXiv:2004.13203*, 2020.

[Pal, 2017] J. Pal. CHI4Good or Good4CHI. In *Proc. of the CHI'17*, 2017.

[Pérez Báez *et al.*, 2019] G. Pérez Báez, R. Vogel, and U. Patolo. Global survey of revitalization efforts: A mixed methods approach to understanding language revitalization practices. *Language Documentation & Conservation*, 13:446–513, 2019.

[Pine *et al.*, 2022] A. Pine, D. Wells, et al. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proc. of ACL'22*, 2022.

[Rangel, 2019] J. Rangel. Challenges for language technologies in critically endangered languages. In *UNESCO LT4All*, Paris, France, December 2019.

[Raval, 2019] N. Raval. An agenda for decolonizing data science. *Spheres: Journal for Digital Cultures*, 5:1–6, 2019.

[Rivenburgh, 2013] N. Rivenburgh. Media framing of complex issues: The case of endangered languages. *Public Understanding of Science*, 22(6):704–717, 2013.

[Sahota, 2007] P. Sahota. Research regulation in American Indian/Alaska native communities: Policy and practice considerations. In *NCAI*, 2007.

[Smith, 1999] L. Smith. *Decolonizing methodologies: research and Indigenous peoples*. Zed Books ; University of Otago Press, 1999.

[Straits *et al.*, 2012] K. Straits, D Bird, E. Tsinajinnie, J. Espinoza, et al. Guiding principles for engaging in research with Native American communities. *UNM Center for Rural and Community Behavioral Health*, 2012.

[Thomason, 2015] S. Thomason. *Endangered languages*. Cambridge University Press, 2015.

[Ubaleht, 2021] I. Ubaleht. Lexeme: the concept of system and the creation of speech corpora for two endangered languages. In *Proc. of ComputEL'21*, 2021.

[van Esch *et al.*, 2019] D. van Esch, B. Foley, and N. San. Future directions in technological support for language documentation. In *Proc. of ComputEL*, volume 1, 2019.

[Wilson, 2008] S. Wilson. *Research is ceremony: Indigenous research methods*. Fernwood Publishing, 2008.

[Wisniewski *et al.*, 2020] G. Wisniewski, S. Guillaume, and A. Michaud. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proc. of Joint Workshop SLTU and CCURL*, 2020.

[Wurm, 2001] S.A. Wurm. *Atlas of the world's languages in danger of disappearing*. Unesco Pub., 2001.

[Xue *et al.*, 2021] L. Xue, N. Constant, A. Roberts, M. Kale, et al. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL'21*, 2021.

[Yangarber, 2018] R. Yangarber. Support for endangered and low-resource languages via e-learning, translation and crowd-sourcing. In *Proc. of FEL*, pages 90–97, 2018.

[Zoph *et al.*, 2016] B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.

[Zuckerman *et al.*, 2021] G. Zuckerman, S. Vigfússon, M. Rayner, N. Chiaráin, et al. LARA in the service of revivalistics and documentary linguistics: Community engagement and endangered languages. In *Proc. of ComputEl'21*, pages 13–23, 2021.