Language and Law / Linguagem e Direito, Vol. *12*(1), 2025, pp. 21-38
DOI: doi.org/10.21747/21833745/lanlaw12_1a1

No language left behind? Towards an integrated framework for linguistic rights, human rights and technology regulation

Ingeborg Birnie [1]

[1] *University Of Strathclyde, Glasgow*, Ingeborg.birnie@strath.ac.uk

Abstract
This paper explores the intersection of artificial intelligence (AI) governance and linguistic diversity, focussing on the digital challenges minority and endangered languages face. It presents a novel framework that applies the lifecycle of digital systems—input, process, and output—to analyse the barriers faced by these languages in AI technologies. Through this lens, it critically examines existing international and European legal instruments, revealing a significant policy gap in addressing the explicit inclusion of minority languages in digital and AI domains. The paper's key contribution lies in its proposal for an integrated, multi-framework approach that combines human rights, minority language protection, and AI governance to ensure equitable linguistic representation. It argues that without urgent and coordinated action, many languages risk digital extinction, with profound implications for cultural identity and access to information. This work calls for comprehensive regulatory reform to secure a digitally inclusive future for all linguistic communities.
Keywords: *Language Rights; AI regulation; minority languages; digital inclusion; linguistic justice.*

Resumo
Este artigo explora a interseção entre a regulamentação da inteligência artifical (IA) e a diversidade linguística, tendo como foco principal os desafios digitais que as línguas minoritárias e as ameaçadas pelo risco de extinção enfrentam. Apresentamos uma estrutura inovadora que aplica o ciclo de vida dos sistemas digitais — entrada, processamento e saída — para analisar as barreiras enfrentadas por essas línguas nas tecnologias de IA. Nesta perspetiva, examinamos criticamente os instrumentos jurídicos internacionais e europeus existentes, onde se revela uma lacuna significativa nas políticas para abordar a inclusão explícita das línguas minoritárias nos domínios digital e de IA.

A principal contribuição deste artigo reside na proposta de uma abordagem integrada e multifacetada que combina direitos humanos, proteção das línguas minoritárias e regulamentação da IA para garantir uma representação linguística equitativa. Discutimos ainda que, sem uma ação urgente e coordenada, muitas línguas correm o risco de extinção digital, com profundas implicações para a identidade cultural e o acesso à informação. Concluímos apelando a uma reforma regulatória abrangente para garantir um futuro digitalmente inclusivo para todas as comunidades linguísticas.
Palavras-chave: *Direitos linguísticos, regulamentação da IA, línguas minoritárias, inclusão digital, justiça linguística.*

1. Introduction

The last few years have seen a significant rise in the public awareness, availability, and use of generative artificial intelligence (AI) tools for a range of different purposes and across different domains. These tools and technologies are designed to interact with the user and are creating seemingly new and meaningful content (Feuerriegel, Hartmann, Janiesch, & Zschech, 2024) which is increasingly difficult to distinguish from responses that might have been created by humans (Mijwil et al., 2023) and therefore blurring the lines between content created in the real and digitally worlds (Ferrara, 2024). These developments have been described by Krishna (2024) as being part of the 'fourth Industrial Revolution' and have the potential to impact different aspects of society, communities and lives of individuals (Farina, Zhdanov, Karimov, & Lavazza, 2024). This has been recognised at an international level by the United Nations who identified that AI technologies and developments can 'improve access to information, health, education, and public services', whilst at the same time acknowledging that these developments in AI have the potential to 'dramatically intensify online harms' (2023).

This recognition of the potential duality of AI technologies has resulted in the United Nations Advisory Body on Artificial Intelligence, who produced their final report in 2024, calling for "a holistic vision for a globally networked, agile and flexible approach to governing AI for humanity … [to] address the multifaceted and evolving challenges and opportunities AI presents … promoting international stability and equitable development" (2024, p. 9) through a flexible framework that balances innovation with safeguards to protect public interest. This recommendation resulted in the adoption of a resolution by the United Nations General Assembly in 2024 that called on Member States to ensure that AI systems operate in compliance with international rights legislation (Mishra, 2024). This resolution, which was universally supported, has come at a time where there have been a variety of initiatives to regulate AI at national, but especially supra-national level. These have typically focussed on the risk that these technologies pose to either individual or state values (Fink, 2021) or are linked to (data) governance (United Nations AI Advisory Body, 2023) but have not explicitly considered the way these AI developments impact on the way individuals use technology to communicate with each other, the world around them, but also the impact on the linguistic and cultural diversity globally.

Linguistic and cultural diversity is under threat globally, with only 1% of the approximately 7,000 languages used in the world today considered to be 'safe' (UNESCO, 2023). The remainder, to a greater or smaller extent, is at risk of disappearing as a community language as a result of both external and internal pressures on the communities that use these languages (UNESCO Ad Hoc Expert Group on Endangered languages, 2003). The extent to which languages are endangered and at risk of disappearing varies. Typically, this is evaluated based on a series of different factors, including the size of the speaker population (both in absolute terms as well as the proportion of speakers vis-à-vis other language communities), the levels of intergenerational transmission, availability of the language in the education system, and also the domains in which the language can be used. One of these domains is the response to new media, including digital spaces, with the recognition that the availability of a language in online domains affect the (wider) perceptions around the functionality and possible uses of the language in contemporary society (Cunliffe, 2007). Additionally, there is an acknowledgement that the absence of languages in new technologies affects the way individuals can interact with these tools, with further recognition that developments in the digital era will significantly impact on how languages are represented in AI applications.

Initiatives to regulate AI at a supra-national level, for example, the EU AI Act (Council of Europe, 2024b) have focussed on the risk that these new technological developments pose to either individual or state values (Fink, 2021). They have also focussed on governance (United Nations AI Advisory Body, 2023), or the wider ethical consideration around the use of AI – including the environmental impact. There has also been some recognition that any measure needs to consider the way individuals use these technologies to communicate with each other, the world around them, including being cognisant of the linguistic and cultural diversity, and that 'no human being or human community should be harmed or subordinated … during the life cycle of the AI system' and that any such technologies should "protect, promote, and respect human rights, fundamental freedoms, human dignity" (UNESCO, 2021, p. 18; paragraph 14).

Furthermore, this UNESCO recommendation on the Ethics of AI recognises that that 'local knowledge, cultural pluralism, value systems and the demands of global fairness to deal with the positive and negative impacts of AI technologies' (UNESCO, 2021, p. 6). However, to date, this recognition has not resulted in specific measures to actively develop the inclusion of linguistic diversity within these AI tools and applications. This article will provide an overview of the current position and role of minority languages in the digital domains: a precursor to their availability and presence in AI tools and technologies. Through an analysis of the life cycles of digital systems, this article looks at the current European policy and legislative frameworks, which aim to support regional and minority languages. It also observes how these can be used to strengthen the position of these languages and ensure equitable access and opportunities in current and future AI developments for the users of these languages. The issues discussed in this article, although framed within the regional and minority language paradigm, act as a proxy for the issues small(er) state languages are facing, and will continue to face, if decisive action to promote inclusive linguistic and cultural practices are not implemented as a matter of priority.

2. Linguistic diversity and digital domains

The presence of languages in digital spaces and tools is an important indicator of their current availability in AI applications and tools. The United Nations resolution *Seizing the Opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* recognised that AI systems need to "advance, protect and preserve linguistic and cultural diversity, taking into account multilingualism in their training data and throughout the life cycle of the artificial intelligence system" (United Nations, 2024). This call to include multilingual training data is particularly pertinent in the context of the current linguistic representation in online space, which forms the basis of the datasets used to underpin the AI technologies.

The most widely used languages in online spaces are mainly those with European origins and spoken (either as a first or, increasingly, additional language) across the world as a result of widespread colonisation. Initially this colonisation involved the physical dominance of some states over others, followed by the subsequent direct or indirect imposition of the language of the colonisers in all or some domains (Liu, 2024; Phillipson & Skutnabb-Kangas, 2017. However, as described by Lukianenko (2024), this process of colonisation is no longer associated with the appropriation of land and infrastructure, but instead, through more subtle means, including through the management of, and access to, digital tools, resources and information. The growth of the internet and digital tools create a new global power dynamic as well as exacerbate the already existing socio-economic divide between different communities.

This digital imperialism exerts power over political, economic, and cultural spheres (Akıner, 2024), and has a limited number of actors. These actors are, unlike more traditional forms of colonialism, not state actors, but internationally operating for-profit organisations, which "at first glance [...] seem to be the providers of unlimited access to information and entertainment and the free circulation of information" (Akıner, 2024, p. 135). They also control the type of information, tools and technologies available through their global reach and dominance. This has been recognised by the United Nations AI Advisory Body (2023) who have acknowledged that access to technology is not universal with "developments and rewards [...] currently concentrated among a small number of private sector actors in an even smaller number of states" (p. 5). The subsequent dominance of a small number of languages in digital spaces becomes clear when the availability and presence of different languages in online domains is analysed: English is the (main) language used in 63% of all websites, with a further 9 languages (beyond English) making up 75% of all the internet content (Spanish, Russian, German, French, Japanese, Portuguese, Turkish, Italian, and Persian) (The Centre for Internet and Society, Oxford Internet Institute, & Whose Knowledge, 2022).

These languages, according to the Centre for Internet and Society et al (2022), "all have either a European colonial history ... or are dominant in specific regions where other languages struggle to remain relevant" (n.p), a view further supported by Jin (2013) who identified that although the mechanism by which this colonisation is happening has changed, the languages that are involved have remained the same. The limited use of languages in online spaces means that to access a wide variety of different types of content, tools and information, individuals will need to learn and then use these in a language in which these materials are available (Pimienta, 2022). This digital imperial-

ism disproportionately affects languages that are considered to be endangered, as the users of these languages have typically already faced assimilation into larger or state languages to access goods, services and information. This includes those provided by the state, and entails challenges to maintain their linguistic and cultural practices, while facing with the further additional barriers in having to learn another (international) language to interact with digital tools and technologies. Research in the European context has shown that lack of availability of resources and tools in a language is a significant barrier to their use. Fewer than 10% of all people in the European Union are "willing or able to use online services in English" (Rehm & Uszkoreit, 2013, p. 22) despite generally high levels of linguistic competences in the language. Not only does this limit individuals' (full) access to digital spaces and technologies, it also further contributes to the digital divide as those with higher levels of education will be more likely to be able to access and use these tools (Cruz-Jesus, Vicente, Bacao, & Oliveira, 2016).

Even though the impact of digital imperialism is, perhaps, most obvious for languages that are currently considered endangered or at risk of disappearing as community languages, this process is beginning to impact languages which, measured by conventional language vitality assessments, would be considered 'safe'. This includes, in the European context, many of the official state languages (Rehm & Uszkoreit, 2013). These state languages are typically being used in less than 1% of all available internet content (the Centre for Internet and Society et al, 2022) with research by Rehm (2018) suggesting that digital support for many of these state languages are non-existent. Unlike minoritised or endangered languages, these state languages might not currently face the pressures of language shift in non-digital domains. However, as Shohamy (2006) identified, the language used online impacts and influences off-line communications. The dominance of a small number of languages in digital spaces is likely to influence long-term linguistic practices.

As already stated, the limited availability of digital tools across different languages is, in part, due to the commercial nature of these technologies: tools and resources are typically developed and promoted by (large) for-profit organisations that operate at an international level (Flew, Martin, & Suzor, 2019), aiming to attract a large user-base (Gurbanova, 2022). Small(er) state, regional, and minority languages typically have a smaller user population, and this means that technological developments in these digitally disadvantaged languages is not as economically attractive. This means that these languages, unless supported at state level or through active requests by the community, will not be (automatically) considered for inclusion. This results in fewer digital resources, tools, and provisions influencing the ability of individuals and communities to "share in scientific advancement and its benefits", as set out in Article 27 of the Universal Declaration of Human Rights (United Nations, 1948).

3. European policy and legislative frameworks

The commercial nature and international reach of digital platforms means that their regulation is complex. Digital platforms typically have their own mechanisms for regulation, although according to Cunningham and Craig (2019) and further supported by Obia (2023), this mainly serves the corporate interests of the platform rather than offering explicit support and protection to its users. These monitoring and moderation procedures can include some mechanisms to support and protect vulnerable groups

(Bruning, Alge, & Lin, 2020). However, their implementation relies mostly on the policies and wider governance instruments and their application by the platforms themselves.

At the state level, there have been requests for content or information to be removed from these platforms (Leerssen, 2015). X (previously known as Twitter) for example, received 97,006 government requests for content to be removed, including 3,831 from the European Union countries in the period July to December 2024 (X, 2025). Furthermore, there have been instances of states implementing restrictions and technical restraints to block or limit access to certain types of applications or content, for instance, in China or in Brazil where social media platforms have been blocked on either a permanent or temporary basis. De Souza Abreu (2018) suggests that this allows states to exert control over the content available to individuals living within (or as a part of) their jurisdiction. These decisions can be based on (alleged) violations of legal and constitutional provisions.

Elements of the digital applications, platforms, and technologies, will be subject to local, regional or state level regulation and support – for example through providing the infrastructure required to access digital technology or through limiting certain types of content and response (AlAshry & Al-Saqaf, 2024; Bao, Sun, & Teplitskiy, 2025). Yet, the nature of the advancement in technology – including the governance of characters and scripts and how these are rendered through input systems (for example keyboards) or displayed on an output system (e.g., a display), means that many aspects require a multi-agency and international approach towards regulation and governance. Many of these regulations - as will be discussed below – do not consider the nature of the language except as being a 'system'. Although, provisions may be made for input to enable the possibility to use minority languages that do not use a standard script or have diacritics to accurately represent the language.

At the European level, support for digital domains and online spaces for small(er) languages and their user community is implicitly included in the two current frameworks that aim to protect linguistic and cultural diversity of (indigenous) minorities: the European Charter for Regional and Minority Languages (ECRML); and the Framework Convention for the Protection of National Minorities (FCNM). It is important to recognise that not all countries, which are part of the Council of Europe, have ratified and / or are signatories to either or both the ECRML with and the FCNM. Furthermore, there is a degree of flexibility in the application of their provisions in a particular state, with both elements affecting the extent of the overall implementation. The ECRML and FCNM fall under the Democracy and Human Dignity Directorate that aims to protect human rights and dignity, strengthen democratic governance, foster innovation, and promote participation and diversity. It also aims to fulfil a separate yet complimentary role in supporting minority communities in the Council of Europe area.

The aim of the ECRML is to protect the historic regional and minority languages (RMLs) – defined in the charter as those languages "traditionally used within a given territory of a State by nationals of that State who form a group numerically smaller than the rest of the State's population and which are different from the official language(s) of that State" (Council of Europe, 1992). The remit of this Framework Convention goes beyond that of ECRML to cover wider issues associated with the rights of persons belonging to national minorities to enable them to "express, preserve and develop [the ethnic, cultural, linguistic and religious] identity [of each person belonging to a na-

tional minority]" (Council of Europe, 1995, p. 1). However, there are also significant areas of overlap between the ECRML and the FCNM in terms of support for the use of minority languages, including its provisions in the media. However, as recognised by Oeter (2013), the ECRML and the FCNM approach minority languages from different angles, with the ECRML setting the standards for positive action, while the FCNM takes a "classical human rights protection" approach (p. 220). Both the ECRML and FCNM also include languages that are used (as main or state languages) within 'kinstates' and thus might receive support through different measures – including at the state level – although this might not necessarily result in a high level of digital inclusion (see Rehm, 2018).

These frameworks aim at protecting small(er) languages and their communities in general, including both implicitly and explicitly supporting their (increased) use in a variety of different domains, for example in interactions with the authorities, the education system and the media. This also (indirectly) includes digital spaces. The aspects of both these policy and legislative frameworks, which most obviously link to the developments in AI and technologies, are the provisions made regarding the inclusion of these languages in the mainstream media. Both the ECRML and the FCNM reference this domain explicitly (Article 11 and Article 9 respectively). However, both the ECRML (and the FCNM) predate the rise of the internet and the increasing influence of online and digital media on the lives of individuals. This includes those affiliating with (national) minorities and how individuals and communities might (potentially) be using their respective languages in these domains to support their overall maintenance and revitalisation. Thus, the ECRML and FCNM focus more explicitly on traditional media (radio and television) which are also (more readily) regulated through traditional measures at state level.

This has resulted in McMonagle (2012) suggesting that technological advancements have been (largely) overlooked "either by states that are party to the ECRML or by the Committee of Experts that conducts three-yearly monitoring exercises of those states and their regional and minority languages" (p. 7). These concerns were brought to the attention of the Committee of Experts of ECRML (COMEX) in December 2024 (Council of Europe, 2024a), following the report produced by Jones, Lainio, Moring, and Resit (2019). With this report called for a mechanism for assessing the use of new media *im* and *for* RMLs in the wider context of the monitoring of this charter, whilst at the same time, acknowledging the challenges of explicitly monitoring the inclusion of regional and minority languages in digital and online domains within the parameters of the current framework of the ECRML.

Although not receiving as much (public) commentary within the academic discourse, similar issues have arisen in the State Opinions of the Advisory Committee of the FCNM that have considered and commented on the availability in online spaces and digital domains of languages used by national minorities as part of the wider media-environment. This exists typically through the availability and use of social media platforms which allows community members to exchange and share information and communicate with each other. This has also included online communication with the authorities at local, regional and national level, for example through the provision of information or service portals in different languages. To date, there have not, as yet, been any active recommendations through the Opinions to States to support the inclusion of

these languages in online spaces. That is, there has not been support beyond measures to make information available to all national minorities and ensure that individuals affiliating with national minorities can participate in the "cultural, social, and economic life and in public affairs" (Article 15).

This does not mean that new tools and applications, including those using AI, have not been considered in the implementation of the various aspects of these frameworks, and in particular the ECRML. Gerken (2022), in her report written with the Secretariat of the ECRML, identified how different tools and technologies could benefit RMLs users and enhance the extent to which these languages can be used to access services (including with the public and judicial authorities), support language acquisition, and increase their use in private domains, a view also supported by Glass, Inge, and Ross (forthcoming). This report also recognised that, although these developments would be highly desirable, 'resolute action' is required to ensure that these technologies can be made available in these languages (p. 21) but without addressing what this action might consist of and how any such recommendations might be enacted in practice.

As identified by Grützner-Zahn and Rehm (2022) any such 'resolute action' to support regional and minority languages (RMLs) in digital spaces should consider the wider socio-economic ecosystem of the language, including the existing support (both financial and in terms of the availability of digital technology). Moreover, these should also aid the capacity from within the language community to support such developments. This is important as technology is increasingly complex, relying on the interplay of many different components that make up the lifecycle of a digital system. These digital systems consist, at their most basic, of three interrelated elements: the input, process, and output. This means that any action to support inclusive practices for minority languages needs to consider these various aspects of the lifecycle (Birnie, Ross, & Glass, 2025): each of these aspects contribute to the overall availability, quality and quantity of digital provisions, and, as will be seen in the following discussion, rely on different policy and regulatory frameworks for their regulation.

4. Life-cycles of digital systems

AI technologies generally require a large training corpus to ensure an accurate response, in terms of the extent of content these tools can produce, as well as the linguistic range and accuracy within a given linguistic context (Le, Bigi, Besacier, & Castelli, 2003). The training data for these technologies is typically based on the publicly available texts and other online resources developed in the digital era. This means that (within the current technological frameworks) the languages that currently have the greatest digital representation will also have the most accurate and advanced outputs, further increasing the already existing digital divide between languages. To establish the challenges and barriers that digitally disadvantaged languages face, including those that are covered by the ECRML and the FCNM, there has to be an evaluation of the current challenges and barriers these languages are experiencing. It also needs to include considerations for the provisions that are required to ensure that they can be represented in AI technologies. This supports the goal, in the words of UNESCO (2021) that "no language is left behind" (n.p.). This analysis needs to consider the basic building blocks that make up online technology cycles: the input, the process, and the output.

Any digital system starts with the input. The input can be defined as the initial interaction by the user of the application that activates the communication between an individual and the technology. This interaction needs, at its most basic, access to technology, incorporating both the physical hardware (through internet enabled devices) and reliable connectivity. Barriers to access include physical connectivity issues (for example, access to broadband or reliable mobile internet access), as well as costs associated with accessing technology hardware (including devices and equipment). Additionally, the digital literacy of individuals within a community also presents barriers. These various factors affect the overall use of digital technologies for all individuals. In particular, minority language users and their communities are more likely to be economically and socially disadvantaged (UNESCO, 2003), and this is also reflected in their overall access to digital tools and access (Duarte, 2017; Steinhauer-Mozejko, 2024).

Minority language users are more likely to live in rural or remote areas where connectivity might be more limited, with a higher reliance on mobile signals or satellite connections (Soylu & Şahin, 2024). Providing this infrastructure in the first instance requires a high level of investment, both at the state as well as international level, and typically relies on partnerships with for-profit organisations. These partnerships make their availability vulnerable to changes in socio-political circumstances and the power relationships between states and for-profit organisations (Abels, 2024), with Gertz and Evers (2020), recognising that "businesses have become key actors in contemporary politics" (p. 199). Furthermore, although these provisions might contribute to reducing the digital divide, their relatively high costs (Oughton, Amaglobeli, & Moszoro, 2023) contribute to, or even increase, the socio-economic pressures these communities might already be facing (Pinhanez, Cavalin, Vasconcelos, & Nogima, 2023; Tepper, 2023).

Even where access to these technologies is available, a further barrier is the interaction with the system to be able to provide an input. Currently, the most common modality used in digital spaces is written text. To allow a language to be used as an input to a digital system the characters that make up the language (including any diacritics or additional symbols) need to be recognised. This requires the language and all its characters to be represented in Unicode – the underlying internationally agreed standards that govern the use of texts in digital systems. It also needs to be recognised and included by input systems (such as keyboards) (Diki-Kidiri, 2009; Yacob, 2006) and can be displayed on a screen (Hossain, 2024). Both the inclusion of characters in the Unicode database and the creation of tools that allow for the creation of inputs, can be initiated by the language community. This can be a time-consuming process, requiring technical expertise that might not be available within the language community itself.

Unicode encoding and support for input tools requires the languages to have an (agreed) orthography that the users are familiar with. This familiarity will, in turn, depend on the literacy levels of the language users. This is particularly pertinent in terms of the input, as this determines the quality and accuracy of the next stages of the digital life-cycle (the process and the output). Minority language users face additional challenges here, as to acquire a high level of literacy in their language there has to be some educational provision to support the development of these skills. Chiaráin et al. (2022) suggest that the level of literacy in a language determines the confidence and willingness of individuals to use their language in digital spaces.

This is very much an under-rated and under-reported factor in discourse around the equitable provision of digital tools for minority languages. Focus is placed typically on the availability of suitable input mechanisms, support for the process and the accuracy and inclusion of these languages in the output. Although education is recognised by both the ECRML and the FCNM (mirroring the Universal Declaration of Human Rights (United Nations, 1948)) as being an important contributor to the knowledge of the language (and culture), especially in communities where intergenerational transmission might be limited, it also plays a fundamental role in creating the conditions for current and future use of languages, across all domains including digital spaces. This is especially the case in contexts where the language might have a largely oral tradition, and therefore literacy is not necessarily part of the community linguistic practices.

With this said, although the majority of inputs will be based on written texts, and thus orthography and literacy skills, increasingly there are different input modalities, which include spoken interactions. Perhaps even more so than the development of input text-based tools, these developments require significant community support and involvement to meet the minimum required input (de Wet et al., 2023). This firmly places the emphasis on individuals and communities themselves rather than being supported through policy or other regulatory frameworks (Armentano-Oller, Marimon, & Villegas, 2024). Different modalities, especially where these contain audio or video materials that can identify the user of the language, are particularly sensitive to data breaches - an issue that has already received attention from international regulatory frameworks (as will be discussed in the next session).

The second element of these digital systems life cycles is the process: the way in which the initial input is analysed and a response formulated, which is then presented as the output. Current AI technologies are founded on the presence of publicly available corpora (typically from internet or other open-sources) to create a large language model (LLM): an AI model that is trained on a large data set (Ozdemir, 2023), which then uses statistical modelling to produce an output (Aydin & Karaarslan, 2023). The output is designed to approximate human-like responses (Feuerriegel et al., 2024). The mechanism by which this typically happens is hidden from the user (a "black box" (Card, 2017)) and involves the use of algorithms, frequency analysis, and statistical inferences. The operation of the black box is not language specific: the processes can be trained on any data, assuming there is a sufficiently large data sample to allow the input to be processed into a meaningful output (Aydin & Karaarslan, 2023). As with all statistical modelling, the larger the corpus, or training data, the more accurate the output produced will be. This is true in terms of the content it can produce,as well as the linguistic range and accuracy within a given language context (Le et al., 2003).

Even where a digitally disadvantaged language (including all the languages covered under either the ECRML and FCNM) is recognised as an input to the system and can be displayed, it might not have a significant online presence. This impacts the size of the corpora available, and thus, the quality of the responses that the system is able to provide – resulting in inaccurate, incorrect, or unreliable outputs. This means that within the current technological frameworks and paradigms languages with the greatest digital representation also have the most reliable and advanced output, contributing to the digital divide (Cahyawijaya, 2024).

Furthermore, in smaller corpora there is a significant risk of data protection legislation breaches, including that of data, which can be directly traced to individuals, locations, or events within the community. As Pinhanez et al. (2023) discussed, digital technologies have played, and are continuing to play, an important role in language documentation initiatives as part of revitalisation efforts. This has, in some instances, included the digitisation of materials recorded and collected in the pre-digital era– including those collected from individuals who are no longer alive (Cahyati & Madya, 2019). Data governance, particularly that which relates to AI technologies, has been a significant concern of the Council of Europe. This has been reflected in their Convention on Artificial Intelligence, the first legally binding international treaty in this field, which opened for signatures in September 2024 (Council of Europe Committee on Artificial Intelligence, 2024). This Act (under Article 11), recognises the need for personal data and privacy rights of individuals to be protected through "applicable domestic and international laws, standards and frameworks" with elements of this covered under the European General Data Protection Regulation (GDPR) framework. The GDPR framework allows individuals to control their personal data in the face of technological advances (Torre, Alferez, Soltana, Sabetzadeh, & Briand, 2021). However, the complexity of the statistical analysis in the "black box" makes it complex to understand and evaluate how personal data is used and then rendered in the output. The likelihood that any data could potentially be used to 'train' these databases might make individuals more reluctant to engage with digital spaces – resulting in an even smaller corpus.

Although these concerns might be related to personal data governance, issues around literacy might also affect the willingness of individuals to create content, but also the quality of the data. In many instances this will relate to written texts, and can also consist of audio or video content and / or other modalities in which languages are used. The quality of the data determines how this can be used in applications – data can be poor in quality because of the input mechanism (for example, where materials are transcribed or transferred from physical sources), or due to their age or limited range of domains. This can also be as a result of errors in the input itself, for instance through spelling mistakes or incorrect grammatical constructions. The quality of the input is significant as any biases in the system will be introduced within the dataset that is used to train the system. In a small dataset, as is typically the case for minority languages, any such errors or limitations are more prone to amplification as there will not be enough additional input to counter these. This will also result in users of these systems and applications becoming more biased and prone to promoting stereotypes (Kotek, Sun, Xiu, Bowler, & Klein, 2024), as they assume that the output generated is accurate (Glickman & Sharot, 2024).

The final element of a digital systems lifecycle is the output or 'end-stage': the combined product of the input and the process. The output is the response that the user receives as a consequence of their interaction with the system. The quality of the output content will be very dependent on the earlier stages – the input as well as the way in which the data has been processed. As explained by Bender, Gebru, McMillan-Major, and Shmitchell (2021), AI systems act as a "stochastic parrot" where the output is created by the "haphazardly stitching together [of] sequences of linguistic forms it (= the system) has observed in its vast training data, according to probabilistic information about how they combine, but without reference to meaning" (p. 617). The quality of

the output is significant as any biases in the system, which in turn, will be introduced within the dataset that is used to train the system, and which will be (negatively) affected by the size, will be amplified. This will also result in the users of these systems and applications becoming more biased, as they assume that the output generated is accurate (Glickman & Sharot, 2024).

Furthermore, the combination of limited data that is used to create a corpus, coupled with a small(er) user base, results in more limited output and thus availability of tools and technologies using these languages. Many projects claiming to be inclusive, for example the "No Language Left Behind" project (META, 2025), supported by UNESCO, only include a limited range of languages, in this case 200 (representing around 3% of all languages), and Wikipedia available in 340 languages. Both examples include minority or endangered languages, and aim to allow individuals to share information and communicate regardless of their language preferences, but have large discrepancies in training materials supporting the creation of the LLM underpinning these tools. They are thus not yet able to create an output that is equally accurate and detailed in all the languages.

Outputs that are not 'fit for purpose' or contain significant errors not only risk being used (in turn) to train the corpora further (through a feedback loop) – but also can result in the wrong information being shared, causing potential harm to the community. The quality of the output is significant, not only in terms of the user experience, but also recognising that "everyone has the right to freedom of opinion and expression … and to seek, receive and impart information and ideas through any media" (United Nations, 1948 Article 19) – and that this needs to be accurate and reliable to ensure equitable access to quality information. However, where these digital outputs are not available, or where they are inaccurate or have a more limited functionality, digitally disadvantaged language users and their communities are 'pushed' into using these applications in languages in which they are available to ensure an equitable end-user experience.

5. Conclusions

The breakdown of the technology life cycle into the input, process, and output allows for an identification and breakdown of the challenges that many languages face, those that are already, to a greater or lesser extent, endangered but also, as shown by Rehm (2018), languages which are currently considered to be 'safe' by conventional language assessment measures. Yet, they are at risk of digital extinction. As discussed in this article, each of these stages introduces and compounds the challenges that users of digitally disadvantaged languages face in accessing technologies in these languages.

While the existing international and European minority language frameworks - such as the ECRML and the FCNM - aim to protect linguistic and cultural diversity, the interpretation of the provisions of these instruments will need to be re-adjusted to consider the current complexities of the provision for a digital presence of the languages. Any further regulations need to ensure that representation of individuals, communities, cultures and languages is accurate and a reflection of current values and lived experiences of the communities. This requires concerted efforts on the part of the authorities to ensure that communities are supported in contributing accurate data to any dataset,

whilst also ensuring that there is an appropriate recourse for removing data that does meet current privacy legislation standards.

Furthermore, the issues affecting digitally disadvantaged languages neither fully sit within the human rights framework, nor solely within the wider frameworks for the protection of minority language nor does the AI governance regulation cover all aspects of the life cycle for these languages. Each of the current policy and legislative frameworks in these areas, both at a global level (for example through the United Nations and UNESCO) and the wider European context (through the Council of Europe), approaches these aspects from a different angle, although with clear overlaps between them. Although, in theory, there are a range of recommendations and protections in place to support the inclusion of digitally disadvantaged languages (and specifically indigenous, regional, and minority languages respectively) and protect the individuals and communities that use these languages, all frameworks are implemented and overseen through different mechanisms. This means that the oversight is currently limited. This therefore requires an overarching (regulatory) framework or body that can consider and incorporate these various elements: one that is cognisant of the current lack of digital inclusion of languages, and which recognises that any language can become digitally endangered if the various aspects that make up the lifecycle of technological systems are not respected and supported.

Acknowledgements

The author wishes to extend her thanks to those involved in Language in the Human Machine Era COST Action, especially those participating in the events organised by Working Group 4 (Language Diversity, Vitality and Endangerment) and Working Group 3 (Language Rights), with the discussions held very much informing the contents of this article with a particular thanks to Dr Maggie Glass (TU Dortmund) and Dr Melody Ross (Universität Duisburg-Essen).

References

Abels, J. (2024). Private infrastructure in geopolitical conflicts: The case of Starlink and the war in Ukraine. *European Journal of International Relations*, *30*(4), 842–866. Retrieved from https://journals.sagepub.com/doi/10.1177/13540661241260653 doi: 10.1177/13540661241260653

Akıner, N. (2024). A theoretical approach to the digital colonialism in the context of media imperialism. *ICONSR 2024*, *131*.

AlAshry, M. S., & Al-Saqaf, W. (2024). Constraints on AI: Arab Journalists' experiences and perceptions of governmental restrictions on ChatGPT. *Journal of Information Technology & Politics*, 1–21. Retrieved from https://www.tandfonline.com/doi/full/10.1080/19331681.2024.2421388 doi: 10.1080/19331681.2024.2421388

Armentano-Oller, C., Marimon, M., & Villegas, M. (2024). Becoming a high-resource language in speech: The catalan case in the common voice corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*

Artificial Intelligence Act. (2024). Retrieved from https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf

Aydin, , & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, *11*(3), 118–134. Retrieved from http://dergipark.org.tr/en/doi/10.21541/apjess.1293702 doi: 10.21541/apjess.1293702

Bao, H., Sun, M., & Teplitskiy, M. (2025). Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. *Quantitative Science Studies*, 1–23. (Version Number: 4)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Virtual Event Canada: ACM. Retrieved from https://dl.acm.org/doi/10.1145/3442188.3445922 doi: 10.1145/3442188.3445922

Birnie, I., Ross, M., & Glass, M. (2025). *Addressing Inequalities Faced by Regional and Minority Languages of Europe in the Human Machine Era. Policy recommendations to mitigate harms and facilitate improved access to artificial intelligence for regional and minority languages users across Europe. LITHME Working Group 4.*

Bruning, P. F., Alge, B. J., & Lin, H.-C. (2020). Social networks and social media: Understanding and managing influence vulnerability in a connected society. *Business Horizons*, *63*(6), 749–761. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0007681320300999 doi: 10.1016/j.bushor.2020.07.007

Cahyati, P., & Madya, S. (2019). Teaching English in primary schools: Benefits and challenges. In *3rd International Conference on Current issues in Education (ICCIE 2018).*

Cahyawijaya, S. (2024). *LLM for Everyone: Representing the Underrepresented in Large Language Models* (Doctoral dissertation). University of Science and Technology, Hong Kong. (Version Number: 1)

Card, D. (2017). The "black box" metaphor in machine learning. *Medium*. Retrieved from https://dallascard.medium.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0

Chiaráin, N. N., Nolan, O., Comtois, M., Robinson-Gunning, N., Berthelsen, H., & Chasaide, A. (2022). Using speech and NLP resources to build an iCALL platform for a minority language, the story of An Scéalaí, the Irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages.*

Council of Europe. (1992). *European Charter for Regional or Minority Languages.* Retrieved from https://rm.coe.int/1680695175

Council of Europe. (1995). *Framework Convention for the Protection of National Minorities.* Retrieved from https://rm.coe.int/168007cdac

Council of Europe. (2024a). *COMEX 80th plenary meeting.* Retrieved from https://www.coe.int/en/web/european-charter-regional-or-minority-languages/-/comex-80th-plenary-meeting

Council of Europe. (2024b). *Framework Convention on artificial intelligence, human rights, democracy and the rule of law.* Retrieved from https://rm.coe.int/1680afae3c

Cruz-Jesus, F., Vicente, M., Bacao, F., & Oliveira, T. (2016). The education-related digital divide: An analysis for the EU-28. *Computers in Human Behavior*, *56*, 72–82. doi: 10.1016/j.chb.2015.11.027

Cunliffe, D. (2007). Minority languages and the Internet: New threats, new opportunities. In *Multilingual Matters* (Vol. 138, p. 133).

Cunningham, S., & Craig, D. (2019). Creator governance in social media entertainment. *Social Media + Society*, *5*(4), 2056305119883428. Retrieved from https://journals.sagepub.com/doi/10.1177/2056305119883428 doi: 10.1177/2056305119883428

De Souza Abreu, J. (2018). Disrupting the disruptive: Making sense of app blocking in Brazil. *Internet Policy Review*, *7*(3). Retrieved from https://policyreview.info/node/928 doi: 10.14763/2018.3.928

de Wet, F., Bukula, A., Karsten, W., Puttkammer, M., Schillack, E., Wierenga, R., & Eiselen, R. (2023, January). Localising the Mozilla Common Voice platform for South Africa's official languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, *4*(01). doi: 10.55492/dhasa.v4i01.4437

Diki-Kidiri, M. (2009). *Securing a place for a language in cyberspace.*

Duarte, M. E. (2017). *Network Sovereignty Building the Internet across Indian Country.* University of Washington Press. Retrieved from http://www.jstor.org/stable/j.ctvcwn92r

Farina, M., Zhdanov, P., Karimov, A., & Lavazza, A. (2024). AI and society: A virtue ethics approach. *AI & SOCIETY*, *39*(3), 1127–1140. Retrieved 2025-10-26, from https://link.springer.com/10.1007/s00146-022-01545-5 doi: 10.1007/s00146-022-01545-5

Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, *7*(1), 549–569. doi: 10.1007/s42001-024-00250-1

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126. Retrieved from https://link.springer.com/10.1007/s12599-023-00834-7 doi: 10.1007/s12599-023-00834-7

Fink, M. (2021). The EU Artificial Intelligence Act and access to justice. *EU Law live*, 1–4.

Flew, T., Martin, F., & Suzor, N. (2019, March). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50. Retrieved from https://intellectdiscover.com/content/journals/10.1386/jdmp.10.1.33_1 doi: 10.1386/jdmp.10.1.33_1

Gertz, G., & Evers, M. M. (2020). Geoeconomic Competition: Will State Capitalism Win? *The Washington Quarterly*, *43*(2), 117–136. doi: 10.1080/0163660X.2020.1770962

Glass, M., Inge, B., & Ross, A. R. (2026). Working Group Four: Language diversity, vitality and endangerment. In D. Sayers, M. Glass, H. Kelly-Holmes, & R. Fuchs (Eds.), *Language in the Human Machine Era: new technologies and the coming transformation of language.* MIT University Press.

Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345–359. Retrieved 2025-10-26, from https://www.nature.com/articles/s41562-024-02077-2 doi: 10.1038/s41562-024-02077-2

Grützner-Zahn, A., & Rehm, G. (2022). Introducing the digital language equality metric: Contextual factors. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference.*

Gurbanova, A. (2022). Problems and prospects for minority languages in the age of industry 4.0. In *The International Symposium on Computer Science, Digital Economy*

and Intelligent Systems.

Hossain, A. (2024). Text standards for the "rest of world": The making of the unicode standard and the OpenType format. *IEEE Annals of the History of Computing*, *46*(1), 20–33. Retrieved from https://ieeexplore.ieee.org/document/10384703/ doi: 10.1109/MAHC.2024.3351948

Jin, D. Y. (2013). The construction of platform imperialism in the globalization era. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, *11*(1), 145–172. doi: 10.31269/triplec.v11i1.458

Jones, E. H. G., Lainio, J., Moring, T., & Resit, F. (2019). *New technologies, new social media and the European Charter for Regional or Minority Languages*. Strasbourg: Council of Europe. Retrieved from https://edoc.coe.int/en/minority-languages/8265-new-technologies-new-social-media-and-the-european-charter-for-regional-or-minority-languages.html

Kotek, H., Sun, D. Q., Xiu, Z., Bowler, M., & Klein, C. (2024). *Protected group bias and stereotypes in Large Language Models.* arXiv. Retrieved from http://arxiv.org/abs/2403.14727 (arXiv:2403.14727 [cs]) doi: 10.48550/arXiv.2403.14727

Krishna, V. V. (2024). AI and contemporary challenges: The good, bad and the scary. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(1). doi: 10.1016/j.joitmc.2023.100178

Le, V. B., Bigi, B., Besacier, L., & Castelli, E. (2003). *Using the Web for fast language model construction in minority languages.* Eurospeech.

Leerssen, P. (2015). Cut out by the middle man: The free speech implications of social network blocking and banning in the EU. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, *6*(2), 99–119. Retrieved from https://www.jipitec.eu/issues/jipitec6-2-2015/4271

Liu, Y. (2024). Linguistic imperialism as a tool in cultural hegemony: Language decline and revitalization of indigenous communities in Canada. *Lecture Notes in Education Psychology and Public Media*, *47*(1), 136–141. doi: 10.54254/2753-7048/47/20240899

Lukianenko, N. (2024). Language and power: Linguistic imperialism. *International Science Journal of Education & Linguistics*, *3*(5), 41–49. doi: 10.46299/j.isjel.20240305.06

McMonagle, S. (2012). The european charter for regional or minority languages: Still relevant in the information age. *JEMIE*, *11*(1).

META. (2025, February). *No Language Left Behind - Driving Inclusion through the power of AI translation.* Retrieved from https://ai.meta.com/research/no-language-left-behind/#

Mijwil, M. M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A.-H., & Bala, I. (2023). ChatGPT and the future of academic integrity in the artificial intelligence era: A new frontier. *Al-Salam Journal for Engineering and Technology*, *2*(2), 116–127. doi: 10.55145/ajest.2023.02.02.015

Mishra, V. (2024). *General Assembly adopts landmark resolution on artificial intelligence.* Retrieved from https://news.un.org/en/story/2024/03/1147831

Obia, V. (2023). Regulatory Annexation: Extending Broadcast Media Regulation to Social Media and Internet Content. *Communication Law and Policy*, *28*(2), 99–123. Retrieved from https://www.tandfonline.com/doi/full/10.1080/10811680.2023

.2206382 doi: 10.1080/10811680.2023.2206382

Oeter, S. (2013). Working with the Language Charter Committee of Experts. In T. H. Malloy & U. Caruso (Eds.), *Minorities, their rights, and the monitoring of the European framework convention for the protection of national minorities: essays in honour of Rainer Hofmann* (pp. 205–227). Martinus Nijhoff Publishers.

Oughton, E. J., Amaglobeli, D., & Moszoro, M. (2023). What would it cost to connect the unconnected? Estimating global universal broadband infrastructure investment. *Telecommunications Policy*, *47*(10). doi: 10.1016/j.telpol.2023.102670

Ozdemir, S. (2023). *Quick start guide to large language models: Strategies and best practices for using ChatGPT and other LLMs.* Addison-Wesley Professional.

Phillipson, R., & Skutnabb-Kangas, T. (2017). Linguistic imperialism and the consequences for language ecology. In *The Routledge handbook of ecolinguistics* (pp. 121–134). Routledge.

Pimienta, D. (2022, June). Resource: Indicators on the presence of languages in internet. In M. Melero, S. Sakti, & C. Soria (Eds.), *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages* (pp. 83–91). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.sigul-1.11/

Pinhanez, C. S., Cavalin, P., Vasconcelos, M., & Nogima, J. (2023). Balancing social Impact, opportunities, and ethical constraints of using AI in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 6174–6182). Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization. Retrieved from https://www.ijcai.org/proceedings/2023/685 doi: 10.24963/ijcai.2023/685

Rehm, G. (2018). The META-NET strategic research agenda for language technology in europe: An extended summary. *Language technologies for a multilingual Europe*, *4*(19).

Rehm, G., & Uszkoreit, H. (2013). *Strategic research agenda for multilingual Europe 2020.*

Soylu, D., & Şahin, A. (2024). The role of AI in supporting indigenous languages. *Sciences*, *2*(4), 11–18.

Steinhauer-Mozejko, P. (2024). *Nêhiyawak Networks: Native Perspectives of Digital Connectivity.*

Tepper, E. (2023). Space Commercialization is Closing the Digital Divide, but Expanding Global Economic Inequality. *Georgetown Journal of International Affairs*, *24*(1), 55–64. Retrieved from https://muse.jhu.edu/article/897701 doi: 10.1353/gia.2023.a897701

The Centre for Internet and Society, Oxford Internet Institute, & Whose Knowledge. (2022). *State of the Internet's Languages Summary Report.* Retrieved from https://internetlanguages.org/en/

Torre, D., Alferez, M., Soltana, G., Sabetzadeh, M., & Briand, L. (2021). Modeling data protection and privacy: Application and experience with GDPR. *Software and Systems Modeling*, *20*(6), 2071–2087. Retrieved from https://link.springer.com/10.1007/s10270-021-00935-5 doi: 10.1007/s10270-021-00935-5

UNESCO. (2003). *Digital Initiatives for Indigenous Languages.* Retrieved from https://www.unesco.org/en/articles/digital-initiatives-indigenous-languages

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). *World Atlas of Languages (Beta Version).* Retrieved from https://en.wal.unesco.org

UNESCO Ad Hoc Expert Group on Endangered languages. (2003). *Language Vitality and Endangerment.* Retrieved from https://ich.unesco.org/doc/src/00120-EN.pdf

United Nations. (1948). *Universal Declaration of Human Rights.* Retrieved from https://www.un.org/en/about-us/universal-declaration-of-human-rights

United Nations. (2023). *UN Global Communications Chief urges AI developers to "put people before profit".* Retrieved from https://www.un.org/en/hate-speech/ai-concerns

United Nations. (2024). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. 87th Session of the United Nations General Assembly.* Retrieved from https://docs.un.org/en/A/78/L.49

United Nations Advisory Body on Artificial Intelligence. (2024). *Governing AI for Humanity: Final report* (Tech. Rep.). Retrieved from https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

United Nations AI Advisory Body. (2023). *Interim report: Governing AI for humanity* (Tech. Rep.). Retrieved from https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

X. (2025). *Global Transparency Report. H2 2024.* (Tech. Rep.). Retrieved from https://transparency.x.com/en/reports/global-reports/2025-transparency-report#government-legal-andlaw-enforcement-requests

Yacob, D. (2006). Unicode for under-resourced languages. *Strategies for developing machine translation for minority languages*, *33*.