Language and Law / Linguagem e Direito, Vol. *12*(1), 2025, pp. 79-99
DOI: doi.org/10.21747/21833745/lanlaw12_1a4

When Machine-generated Mistranslation on Social Media Becomes Misinformation: Risks to Users, Corporate Responsibility, and Legal Implications

Khetam Al Sharou [1]

[1] *Dublin City University & University of Damascus*, khetam.alsharou@dcu.ie

Abstract
Machine-generated mistranslations on social media can result in misinformation, with potential major consequences for users, especially marginalised communities. As Machine Translation (MT) is increasingly used to access online content, its errors often go unnoticed by users lacking knowledge of the source language. MT inaccuracies can distort meaning, contribute to misinformation, and reinforce digital inequality. Social media has become a main source of information. The unchecked use of machine-generated content introduces vulnerabilities, especially in politically and culturally-sensitive contexts. Through real-world case studies and empirical analysis, this work shows how mistranslations can distort meaning and cause misinformation. It highlights the ethical responsibility of tech companies and service providers to ensure accuracy and transparency while mitigating the risks that arise when MT errors lead to real-world harm. It further assesses how regulatory frameworks, including the EU's Digital Services Act and other similar frameworks, can help address these challenges. This work advocates for responsible MT integration, equitable information access, and stronger corporate and regulatory accountability in combating MT-driven misinformation.
Keywords: *Machine Translation, social media, misinformation, language rights, user experience and corporate responsibility.*

Resumo
Os erros de tradução automática (TA) nas redes sociais podem resultar em desinformação, com consequências potencialmente graves para os utilizadores, especialmente para comunidades marginalizadas. À medida que a TA é cada vez mais utilizada para aceder a conteúdos online, os seus erros passam frequentemente despercebidos aos utilizadores que não têm conhecimento da língua de origem.

As imprecisões da TA podem distorcer o significado, contribuir para a desinformação e reforçar a desigualdade digital. As redes sociais tornaram-se numa das principais fontes de informação. A utilização descontrolada de conteúdos gerados por máquinas potencia algumas vulnerabilidades, especialmente em contextos politicamente e culturalmente sensíveis. Através de estudos de caso reais e análises empíricas, este trabalho mostra como os erros de tradução podem distorcer o significado e gerar desinformação. Destacamos a responsabilidade ética das empresas de tecnologia e dos prestadores de serviços em garantir a precisão e a transparência, mitigando os riscos que surgem quando os erros de TA causam danos no mundo real. Além disso, avaliamos como os quadros regulamentares, incluindo a Lei dos Serviços Digitais da UE e outros quadros semelhantes, podem ajudar a enfrentar esses desafios. Este trabalho defende a integração responsável da TA, o acesso equitativo à informação e uma maior responsabilidade corporativa e regulamentar no combate à desinformação impulsionada pela TA.
Palavras-chave: *Tradução Automática, redes sociais, desinformação, direitos linguísticos, experiência do utilizador e responsabilidade corporativa.*

1. Introduction

False or misleading information, including both misinformation and disinformation, has always been a societal concern, significantly exacerbated by the widespread use of social media platforms as vehicles to disseminate information and misinformation/disinformation. The term misinformation/disinformation began to increasingly appear in scholarly discussions of social media following the 2016 US election, referring to information that is untrue and can mislead people (Comito, 2023, 2024). UNESCO differentiates between misinformation as "information that is false but not created with the intention of causing harm" and disinformation as "information that is false and deliberately created to harm a person, social group, organisation or country" (Ireton & Posetti, 2018). The main concern is the ease with which information, verified or otherwise, is posted, shared and consumed on social media platforms. Statistics indicate that, as of February 2025, 5.24 billion people (63.9% of the global population) were using social media (Petrosyan, 2025). Misinformation/disinformation on social media is thought to have influenced major events worldwide, including those related to health and politics (see, Allcott & Gentzkow 2017; Baker, 2022; Martin et al., 2022)

International organisations have taken steps to raise awareness of the harm that such activities can cause. For example, UNESCO has pointed out that misinformation/-disinformation can spread at a large scale in different shapes and forms, fuelled by new technology especially social media platforms (Ireton & Posetti, 2018). Governments have also introduced emergency laws to prevent the spread of false information online. Regulations to reduce online harms and ensure platform accountability, for example: the Digital Services Act (DSA) - Regulation (EU) 2022/2065 (European Union, 2022); the EU 2022 Code of Practice on Disinformation (European Commission, n.d.); the UK Online

Safety Act 2023 (Department for Science, Innovation & Technology, 2025); and Ireland's Online Safety and Media Regulation Act 2022 (Department of Culture, Communications & Sport, 2022), represent important steps in online content moderation and regulation.

The EU's DSA, fully applicable since February 2024, represents a significant effort to regulate online platforms by imposing stricter transparency and accountability measures, particularly on what it terms as Very Large Online Platforms (VLOPs), which have more than 45 million monthly users within the EU, such as Facebook, Instagram, X, YouTube and LinkedIn. Both the UK's Online Safety Act 2023 (OSA) and Ireland's Online Safety and Media Regulation Act 2022 aim to regulate online activities and address illegal and harmful content. At the EU level, the Code of Practice on Disinformation, revised in 2022, was recognised as a Code of Conduct under the DSA in 2025, transitioning from voluntary compliance to enforceable standards and reinforcing the EU's commitment to establishing a more transparent, accountable, and trustworthy digital information environment (European Commission, n.d.).

These frameworks, while important, primarily focus on the source text, overlooking Machine Translation[1](MT) and related language technologies as potential contributors of information distortion (Al Sharou & Moorkens, 2024). Access to information is recognised as a human right in both international and national legal frameworks. It is protected under Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which affirms every individual's right to seek, receive, and share information (UN, 2022). This right is considered essential for societal participation and for promoting equality (Nurminen & Koponen, 2020). In today's globalised world, access to multilingual information is often supported by advanced technology. Many social media giants now have MT as an additional service on their platforms; for example, Meta has developed its own automated translation systems (models) to increase user engagement and make information available across several languages on its platforms (Costa-jussà et al., 2022; Vincent, 2022).

While MT made access to information in other languages possible, its outputs, when not reviewed by humans, can have critical errors with major impact on the end-users. This particularly holds true for speakers of non-dominant languages who rely on MT to access information or engage in online discussions (see Al Sharou & Specia, 2022; Lee & Qian, 2022; Martindale, 2024, among others). Current MT systems are capable of producing fluent content and some users seem to accept it as correct even when it is not (Lee & Qian, 2022; Martindale, 2024). Therefore, it is crucial when deploying these MT systems to consider language rights, as per the ICCPR, and the risks to these vulnerable communities. These communities, especially those who are not fluent in the original language of a text (or do not have access to translation professionals) are particularly vulnerable to the impact of such mistranslation (Martindale, 2024). This is particularly important as liability for harm caused by MT mistranslations is still an unresolved issue, leaving users in a legally ambiguous position (Canfora & Ottmann, 2020). Therefore, the ethical responsibility of MT providers should extend beyond merely fixing technical

[1]In this case study, we refer to Machine Translation (MT) as an integrated feature within social media platforms, either developed by the platform providers or as a third-party service. The discussion may also apply to other automatic translation systems, including tasks performed by large language models (LLMs), and those offered as standalone tools and apps, available as free-to-use services (paid services are not included).

issues to include proactive safeguards which ensure equal and fair treatment for all users. Without these principles in practice, the goal of inclusivity on global platforms remains unfulfilled.

Translation as a human activity has been tied to issues including: ideology manipulation in translation; how ideology can affect translators' linguistic choices and the impact that it has on the receivers of the translations; and how translations can shape receivers' views of the world (Calzada-Pérez, 2014; Farhan, 2017; Wang & Feng, 2018). For example, bias in news reporting can be displayed through the use of certain stylistic features that can influence how meaning is constructed, forming people's understanding of other cultures and places (see discussion in Riggs, 2019a, 2019b). While MT is not a human activity, its outputs are derived from human communications and can still carry the political biases or culturally preconceived ideas of humans. MT systems are trained on data and algorithms that are created and developed by humans, and research has revealed that the data used to train these machines contain and reproduce bias (see Vanmassenhove et al., 2021). These biases can subtly or overtly influence the output of the system. For example, instances of gender bias, such as the over-representation of masculine pronouns, or even racism, have been observed (Fan et al., 2019; Ghosh & Caliskan, 2023). This study will demonstrate how linguistic bias can be embedded into automated systems in ways that cause real harm, leading to the marginalisation and dehumanisation of entire groups of people. It highlights a broader concern that, without adequate oversight, these technologies can replicate and even amplify the prejudices of the societies that design them.

In this work, we look at MT-driven misinformation from the perspective of how it changes the *inferred intent* and sentiment of the original text. This results in the reader receiving a distorted message that could trigger a certain action or change their opinion about a place, person, action, or event. The current study tackles the challenge of disseminating distorted information through MT by directly utilising real-life examples of mistranslations, and as such, will address social media and MT providers' policy goals on misinformation, automation, and information sharing. The research seeks to highlight the risks associated with these technologies and discuss the ethical and social implications, along with corporate responsibility regarding their development and use. This work answers the following research questions (RQs):

> **RQ1**. To what extend a machine can distort facts, fuel misinformation and reflect inherent biases?
> **RQ2** What are the potential consequences on the end-user and service providers?
> **RQ2.1** Do these consequences, if any, vary depending on the political and cultural contexts, the affected groups and status of the users?
> **RQ3.** What are the ethical implications of deploying and using MT systems in sensitive contexts?
> **RQ4.** What obligations do tech companies and service providers have in mitigating harms, ensuring transparency, and safeguarding users from machine-generated misinformation?

To fulfil the objectives of this study and answer these RQs, four real-life case studies of mistranslations, generated by the auto-translate feature on Meta's Facebook and Instagram platforms, are analysed in depth. It may be argued that reported cases of

mistranslation in social media might be considered as rare, meaning that the **likelihood** of a mistranslation problem causing severe consequences is low given the amount of machine-translated social media content that is consumed on a regular basis. However, these severe consequences do occur, and cannot be ignored, making it an issue deserving of our attention. In this study, misinformation can result from low MT performance, where the machine may, or may not, have necessarily been programmed to produce incorrect information. It is an attempt to highlight the issue and a call for more action to ensure a safer online experience for all users

First, we provide an overview of prior research on the relationship between social media, misinformation, and automated translation. Second, we explain Meta's policy on misinformation and automated translation. Third, we present and analyse four case studies of mistranslations caused by the auto-translate feature on Meta's Facebook and Instagram, discussing its impact on users and the ethical considerations associated with their use and development. The analysis highlights the responsibility of social media platforms to ensure the accuracy of its automated translation systems and explores how existing regulatory frameworks can support equal access to information and the protection of language rights in the digital age. The paper concludes with recommendations for future action and proposes directions for further research.

2. Social Media, Misinformation and Machine Translation

Research has discussed how the use of social media could influence users' political and social attitudes, resulting in the adoption of false or misleading ideas that can sometimes lead to tangible consequences in both behaviour and political discourse (Martin et al., 2019, Pedace; 2021; Saaida, 2023). For example, by investigating findings from multiple sources, Allcott and Gentzkow (2017)'s study demonstrates how misinformation played a significant role in shaping political outcomes during the 2016 US election, benefiting Donald Trump. Their research reveals that false stories favouring Trump's campaign were shared more widely than factual ones, and that many users believed the misleading content they came across (see also Martin et al., 2019). Baker (2022) indicates that social media influenced users' attitudes during Covid by allowing wellness influencers who promote alternative health practices to spread misinformation by presenting themselves as relatable and authentic. The platforms enabled them to challenge mainstream institutions such as governmental research and public health institutions while promoting alternative beliefs and conspiracy theories, encouraging distrust in science and resistance to public health measures (Baker, 2022).

Social media is a communication environment that is extremely diverse. With automatic translation in the form of MT now being very common on these platforms, users consume media produced in different places and in a range of languages. Social media platforms provide raw, unchecked MT output, leaving users with possibly distorted translations, raising ethical concerns due to their potential to generate inaccurate information (Al Sharou & Specia, 2022). Mistranslations by MT can even introduce toxicity or offensive language that was not in the original text, as shown in Al Sharou and Specia (2022). In their study, they looked into the use of MT for user-generated content and the type of errors MT can generate when dealing with such content. Their examples show how some human-generated errors in the content - such as grammatical or orthographical errors - and writing techniques like shortening, word lengthening or dis-

guising words using symbols, can lead to inaccurate translations (see also Al Sharou et al., 2021's study that presents a taxonomy of such non-standard features of the text). These translations may cause misunderstandings or misinterpretations for users who rely entirely on the translation to access the meaning or message of the original content. For example, the sentence "your killing the f***ing planet" in English was translated by Google Translate into Arabic as "may the damn planet kill you", reversing the sentiment and introducing violence (Al Sharou & Specia, 2022). Users can face serious consequences if they believe and act on these inaccurate translations (Al Sharou and Specia, 2022). However, some of these errors may even be difficult to detect, presenting significant risks to users (Canfora & Ottmann, 2020; see also Lee & Qian, 2022). This difficulty in detecting errors is a critical liability concern especially considering that users who are not fluent in the source language can be at higher risk of being misled by mistranslations (Lee & Qian, 2022; Martindale, 2024).

Nevertheless, there is little research on evidence that MT spreads misinformation. To date, only two studies focus on MT-driven misinformation. Narayanan's study examines the spread of misinformation on Facebook due to mistranslations of news headlines from English to Tamil (2022). For example, the English headline, "Trump For Rushing To Defend Tomi Lahren While Ignoring Real Victims," was translated into Tamil as "Trump was for coming early to protect Tomi Lahren ignoring the real victims." The translated version changed the meaning, leading to a misleading interpretation that suggested Trump's actions were more neutral or positive than the original intended critical tone. The study found that twenty percent of general and ambiguous translated headlines, and thirty percent of sarcastic and domain-specific headlines, did not accurately reflect the meaning of the original source. Lee and Qian's (2022) study examines MT-driven misinformation by analysing four English-to-Chinese MT texts, finding that most misinformation was caused by polysemy/named-entity errors and non-equivalent idiomatic expressions. One example from Lee and Qian (2022, pp.538-539), categorised as a semantic error, shows how MT can seriously change the meaning of a sentence: the original English sentence, "…but that we cannot burden people with a carbon tax or a gasoline tax to slow global warming," was mistranslated into Chinese as "…but that we cannot impose a carbon tax or a gasoline tax to lessen people's burden to slow global warming." This completely reverses the intended meaning, creating misleading information.

3. Meta's policy on Misinformation and Machine Translation

Meta is the parent company of Facebook and Instagram, two of the most popular social media platforms globally, with billions of active users interacting and sharing content (see, Meta, n.d.a). Meta's auto-translate feature supports this global communication in several languages. Facebook automatically translates posts based on the user's default language settings (Facebook, n.d). For example, if a user has Arabic set as their default language on Facebook, posts in English will be automatically displayed in Arabic. The auto-translate feature is also available for comments. Furthermore, users have the option to view the content in its original language by clicking on the "See Original" option. Users can rate the translation, but no option to report incorrect translations is given. Instagram provides a "See Translation" option to translate a post's caption, comments, and profiles (Instagram, n.d.). In terms of its effort to develop their MT systems and allow people to communicate in their native languages across all its platforms, Meta launched

the "No Language Left Behind (NLLB)" initiative (Costa-jussà et al., 2022; Meta, n.d.b). By utilising cutting-edge modelling techniques, the project aimed to achieve high translation accuracy and make MT accessible for major languages as well as low-resource languages, including dialects. As of its latest report, NLLB covers over 200 languages with 150 low-resource languages included (see, Figure 1)[2].

Real-World Applications		Experience the Tech	NLLB Innovations		Research Milestones
Arabic (Iraqi/Mesopotamian)	Welsh	Italian	North Azerbaijani	Finnish	Kyrgyz
Arabic (Yemen)	Danish	Javanese	Bashkir	Fon	Kimbundu
Arabic (Tunisia)	German	Japanese	Bambara	Scottish Gaelic	Konga
Afrikaans	French	Kabyle	Balinese	Irish	Korean
Arabic (Jordan)	Friulian	Kachin | Jinghpo	Belarusian	Galician	Kurdish (Kurmanji)
Akan	Fulfulde	Kamba	Bemba	Guarani	Lao
Amharic	Dinka(Rek)	Kannada	Bengali	Gujarati	Latvian (Standard)
Arabic (Lebanon)	Dyula	Kashmiri (Arabic script)	Bhojpuri	Haitian Creole	Ligurian
Arabic (MSA)	Dzongkha	Kashmiri (Devanagari script)	Banjar (Latin script)	Hausa	Limburgish
Arabic (Modern Standard Arabic)	Greek	Georgian	Tibetan	Hebrew	Lingala
Arabic (Saudi Arabia)	English	Kanuri (Arabic script)	Bosnian	Hindi	Lithuanian
Arabic (Morocco)	Esperanto	Kanuri (Latin script)	Buginese	Chhattisgarhi	Lombard
Arabic (Egypt)	Estonian	Kazakh	Bulgarian	Croatian	Latgalian
Assamese	Basque	Kabiye	Catalan	Hugarian	Luxembourgish
Asturian	Ewe	Thai	Cebuano	Armenian	Luba-Kasai
Awadhi	Faroese	Khmer	Czech	Igobo	Ganda
Aymara	Iranian Persian	Kikuyu	Chokwe	Ilocano	Dholuo
			Central Kurdish	Indonesian	Mizo

Figure 1. Full list of supported languages by Meta (Meta, n.d.b)

According to Meta's Code of Conduct for Virtual Experiences (Meta, n.d.c) and Hateful Conduct (Meta, n.d.d), its mission is to empower individuals and give them the opportunity to express themselves freely by creating an online communication environment that is welcoming, non-intimidating, and that does not encourage offline harm or violence. Meta's Community Standards define what is and is not permitted on its various platforms, including Facebook and Instagram (Meta, n.d.e). Meta claims that these standards "apply to everyone, all around the world and to all types of content, including AI-generated content" (Meta, n.d.e).

Meta's misinformation policy aims to provide flexible guidelines to manage false content while balancing free speech with the need to prevent harm (Meta, n.d.f). It recognises that truth is not static but can evolve, and that misinformation is often context-dependent; accordingly, it avoids a blanket ban, acknowledging the nuanced challenges of verification and enforcement (Meta, n.d.f). In 2025, however, Meta stopped its Third-Party Fact-Checking Program in the US, which relied on external organisations to verify content, replacing it with Community Notes, a user-driven system where individuals attach contextual notes to potentially misleading posts (Kaplan, 2025). The move was framed by CEO Mark Zuckerberg as reaffirming Meta's "commitment to free expression" (Kaplan, 2025). However, analysts cautioned that without professional fact-checking, hate speech and disinformation could go undetected, affecting marginalised communities (e.g., ethnic, religious groups), who already faced disproportionate online targeting (Booth, 2025). The European Commission also voiced concern over Meta's de-

[2]Please note that Figure 1 has Modern Standard Arabic (MSA) mentioned twice. It was not possible to confirm with Meta whether it is a mistake or they refer to different variants of Arabic language.

cision, stressing that any content moderation system implemented within the EU must undergo a formal risk assessment, demonstrate its effectiveness, and comply with the Digital Services Act (Tsimitakis, 2025).

Nonetheless, Meta's policies do not include explicit guidelines directed at users specifically on how to report cases of misinformation, generated by its auto-translate feature, nor are there warnings about its tool's limitations. The motivation behind this study is the real-life examples of mistranslations that were caused by the auto-translate feature on Facebook and Instagram. These mistranslations demonstrate how Meta's auto-translate feature can generate errors that are critical, and can significantly distort the intended meaning of content, sometimes with serious and harmful consequences for users.

4. Misinformation Through Mistranslation: Real-life Examples

Four real-life case studies of mistranslations, produced by the auto-translate feature on Meta's Facebook and Instagram platforms, were selected and analysed in depth. We focused specifically on cases that:

- Drew significant attention (reported by major news outlets),
- Directly affected communities (e.g., causing political, social, or cultural harm), and
- Prompted a public reaction from Meta (e.g., corrections, official apologies).

These examples will be discussed in terms of four topics (issues), providing answers to the four aforementioned RQs:

1. **Accuracy of Translation (RQ1)**: Exploring how the MT system misinterpreted the original message and the extent to which the translation deviated from the intended meaning.
2. **Impact on Users (RQ2 & RQ2.1)**: Covering consequences of the errors, such as embarrassment, violence, and/or harm, on the end-users.
3. **Ethical Considerations (RQ3)**: Including ethical implications of deploying and using MT systems in sensitive contexts, showing how the content, and/or the sensitive nature of the context, has made the mistranslation particularly dangerous and damaging.
4. **Corporate Responsibility (RQ4)**: looking at the role of companies such as Meta in ensuring the accuracy and reliability of their MT systems and their responsibility towards addressing issues when they arise, implementing safeguards to protect users from harm caused by incorrect translations.

First, the four case studies will be presented briefly before embarking on discussing them in light of these four topics.

4.1. Example One (EX1)

In a recent incident in 2023, some Palestinian Instagram users' profiles were incorrectly translated, with the term "terrorist" being inserted into their bios. This issue affected profiles that contained the word "Palestinian," the Palestinian flag emoji, and the Arabic phrase "alhamdulillah" ("Praise be to God"). When clicking "See Translation", the English translation read as: "Praise be to God, Palestinian terrorists are fighting for

their freedom" (McMahon & Tidy, 2023). As reported by 404media, the user who initially posted about the issue on Tiktok tried to translate the phrase ("Praise be to God") alone without including the word Palestinian or the flag emoji. The result was a correct translation, "Thank God", that did not include the word "terrorist" (Cole, 2023). Instagram's auto translate feature not only provided a distorted translation but also attached a politically-charged label "terrorist" to a group of people. This led to widespread backlash from users about the platform's biases. The issue was fixed by Meta, who attributed this error to a "technical bug" and apologised (Taylor, 2023). However, critics called for greater clarity around how its translation system operates. As reported by the Guardian, Fahad Ali, the secretary of Electronic Frontiers Australia and a Palestinian based in Sydney, said:

> There is a real concern about these digital biases creeping in and we need to know where that is stemming from. Is it stemming from the level of automation? Is it stemming from an issue with a training set? Is it stemming from the human factor in these tools? There is no clarity on that. (Taylor, 2023)

4.2. Example Two (EX2)

The Guardian reported that a Palestinian man was wrongly arrested due to a critical error in a machine-translated post in Facebook, which led to a belief that he was planning an attack (Hern, 2017). The man, a construction worker in the West Bank, posted "yusbihukum" on his Facebook profile, which translates to "good morning." However, Facebook's MT system mistakenly translated this phrase into "hurt them" in English and "attack them" in Hebrew. The man was detained for several hours, and only after questioning did the police release him. Notably, the post had not even been reviewed by any Arabic-speaking officer before action being taken based purely on the machine-generated translation (Hern, 2017). This is concerning and shows that MT systems can inadvertently create hostility and aggression. The BBC (2017) mentioned that the post was deleted without adding any further details as to why. Nor did the article reflect on the user's experience of the incident. In sensitive environments, the ability of MT systems to distort meaning and generate potentially harmful content is a significant issue. Facebook apologised for the mistake, stating that it was working to address the issue, though it also acknowledged that such errors can occur, even with improvements to their MT systems.

4.3. Example Three (EX3)

In 2020, Facebook's auto-translate feature mistranslated Chinese President Xi Jinping's name from Burmese to English as "Mr. Shithole", an offensive and inappropriate term that was described by the media as "embarrassing" (Serrano, 2020; The BBC, 2020). The mistranslation appeared in Facebook posts shared on the official Facebook page of Myanmar's State Counsellor, Aung San Suu Kyi (McPherson, 2020; The BBC, 2020). In response, Facebook issued an apology, attributing the mistake to a "technical issue" (Serrano, 2020). A spokesman for Facebook said: "this should not have happened and we are taking steps to ensure it doesn't happen again," and "We sincerely apologize for the offense this has caused" (McPherson, 2020). Facebook admitted that the Burmese-into-English database on which the MT system was trained did not contain Xi Jinping's name and that led to the mistranslation (McPherson, 2020; Serrano, 2020). After that,

the English translation function did not appear to be working on the Burmese posts of official Facebook pages belonging to Ms. Suu Kyi and the Myanmar government, suggesting ongoing technical issues (McPherson, 2020; The BBC, 2020). This controversy highlights the political and technical limitations of MT, especially when it comes to incidents involving political figures and cultural sensitivities, leading to potentially harmful consequences.

4.4. Example Four (EX4)

In 2020, Facebook faced another significant criticism when its auto-translate feature mistranslated a headline about the live broadcast of the ceremony celebrating the King of Thailand's birthday posted on the Facebook page of the Thai Public Broadcasting Service (Thai PBS) (Marking, 2020). The original English headline, which was meant to convey respect and celebration, was automatically turned into an offensive phrase. "King's birthday" was changed to "King's Memorial Day," in the Thai translation (Soponvijit, 2022), a term usually associated with mourning someone's death. The mistranslated headline caused a public outcry and serious legal and social repercussions in Thailand, with some viewers demanding the resignation of Thai PBS executives (Soponvijit, 2022). The Royal Thai Police initiated a formal investigation into the mistranslation incident following an official complaint submitted by the Thai PBS to the cybercrime division (Marking, 2020). This example highlights the legal responsibility platforms bear when a mistranslation results in spreading the wrong information. Facebook issued a formal apology, and temporarily deactivated its auto-translate feature on both Facebook and Instagram, promising to enhance the quality of their translation (Marking, 2020). This case also underscores the significant impact that translation errors can have on both legal liability and public trust. According to Marking (2020), Thailand's strict *lèse-majesté* laws make this case particularly significant, as even unintentional mistranslations can lead to imprisonment. Moreover, the fact that the Thai press avoided publishing the exact mistranslation further illustrates the severity and sensitivity of the issue (Marking, 2020). This aligns with the argument that MT errors can have serious real-world impacts, especially in high-stakes situations.

The following discussion examines the central issues brought to light by these examples, namely: the accuracy of MT, its potential impact on users, ethical considerations, and corporate responsibility. It offers a deeper analysis of how machine-generated mistranslations can influence and misinform end-users and the broader implications of such errors.

5. Discussion - Accuracy of Machine Translation (MT)

The presented examples of mistranslations highlight the risks of relying on MT in politically and socially sensitive contexts. They show how MT has the potential to mislead and produce translations that misrepresent the original message and can even criminalise individuals. The mistranslations occurred within the context of user-generated content, which is often more challenging for MT systems to handle. User-generated content is informal, frequently uses slang, and can include context-specific language that automated systems often struggle to interpret correctly (Al Sharou et al., 2021). The broader implication is that automated systems, which may work well in some contexts, still fail to comprehend the complexities of real-world communication and are not

fully ready to deal with such content and context (see discussion in Al Sharou & Specia, 2022).

In some cases (EX1 and EX2), mistranslations include biases and misconceptions that could be the result of training MT systems on biased data (Fan et al., 2019; Vanmassenhove et al., 2021). Some Arabic words related to Islam, such as "alhamdulillah", "Allahu Akbar" and others, are often misinterpreted, leading to misunderstanding and misrepresentation. "Alhamdulillah," meaning "Praise be to God," is a common expression of gratitude and appreciation. "Allahu Akbar" translated as "God is the Greatest," is a phrase used in worship to express devotion and admiration. The media has played a role in distorting their meaning by associating Muslims with violence and terrorism (Riggs, 2019). According to Corbin (2017), the US media often portray Muslims as terrorists with their prevalent biased narrative that "all terrorists are Muslims" which leads to the belief that "all Muslims are terrorists" (p. 457). The study highlights how post-9/11 films often present Arabs and Muslims negatively, and how news outlets such as Fox News during the Quebec City mosque attack in 2017 quickly linked Muslims to terrorism. The issue is further compounded by the predominant framing of Palestinians in public discourse, i.e. the dehumanisation of Palestinians, in particular being described collectively as "terrorist", which has intensified significantly during the current war on Gaza, according to many media studies and news reports (see Khurma, 2024; Osama, 2025). This misrepresentation does not occur in isolation; instead, it echoes and reinforces pre-existing prejudices against the Palestinians. It is plausible that this, along with similar content, has been used to train existing MT systems, thereby allowing dominant narratives to influence and shape the output (see discussion in Nee et al., 2021). For example, in 2022, Google Translate faced criticism after its translation of the Arabic word "takhteet", meaning "to plan," included an example sentence, "planning to blow up the car" (Clark, 2022; Warner, 2022). Google acknowledged the issue, explaining that the error was a result of biased data used to train the MT system and removed the offensive example (Clark, 2022; Warner, 2022). When training MT systems for Arabic language, it is important to approach religious and cultural terms with an awareness of their true meanings and the contexts in which they are used to avoid biases due to selective or inaccurate translations. The issue of terminology in MT is discussed in Canfora and Ottmann (2020), who emphasise the importance of improving both the accuracy of MT and the consistency of terminology to enhance the quality of MT systems, particularly in contexts where specific meaning is critical.

Another significant challenge for MT is the lack of sufficient training data for low-resource languages. MT models depend on large amounts of parallel texts (source texts and their translations) to learn accurate language mappings. However, low-resource languages often have limited digital content, resulting in less accurate and less robust models (Bender et al., 2021; Nee et al., 2021). Both Thai and Burmese are classified as low-resource languages in the context of MT due to the limited availability of high-quality, parallel corpora for training translation models (Tzoneva, 2023; San et al., 2024). This limits the effectiveness and accuracy of MT models trained for these languages. Arabic dialects are also considered low-resource languages because they have limited parallel corpora and lack standardised written forms, especially when compared to Modern Standard Arabic (MSA). This lack of resources creates translation challenges, par-

ticularly when translating between dialects and languages that carry distinct cultural or contextual meanings (Slim & Melouah, 2024).

5.1. Potential Impact on Users

The first two case studies offer an illustration of how MT errors can result in serious and immediate consequences for individuals and can lead to dangerous and unjust actions. They display distinct yet interconnected harms arising from unverified MT. EX1 shows broad harm caused by MT, where Palestinian users were collectively misrepresented through the insertion of the term "terrorist" into their Instagram bios. This critical error led to a distorted and harmful narrative, falsely associating an entire group with terrorism. The effects of such a translation are substantial, as it could incite mistrust, fear, or even aggression towards specific individuals on Instagram, incorrectly presented as a threat. Furthermore, people who read or interact with such output may start to believe those biased views are accurate (Bender et al., 2021). This is especially concerning, as research shows that even when false or misleading information is corrected, it can still influence what people think and believe unless the correction is highly credible and clearly communicated (see discussion in Sanna & Lagnado, 2025). In this case, Meta should have taken further, more effective, steps to address the issue and reduce the potential for misunderstanding and harm. The company could have pinned clarifications to the affected profiles, publicly explained the mistranslation, and outlined concrete measures to prevent future occurrences. It might be argued that such measures are not technologically feasible, but service providers still have a responsibility to act properly when the impact on users is significant.

Furthermore, the deployment of automated systems without adequate quality control has serious repercussions for the reputation as well as the safety of individuals. EX2 shows that such mistranslations are particularly concerning because the fluency and grammatical accuracy of such outputs can mask the fact that they are mistranslations. Such critical errors can lead to misinformation and real-world harm, as evidenced by the wrongful arrest of the Palestinian worker who posted "good morning" on his Facebook page. It is important to note that this unfortunate incident could likely have been avoided if the authorities/security forces had a policy requiring human validation of machine-translated content before taking critical action based on these translations. As Martindale (2020) suggests, institutions relying on MT for high-stakes decisions should ensure verification by someone fluent in the source language. It can be added that the specific political and social context could have also contributed to such an unfair action against the Palestinian worker, underscoring the need for human supervision when deploying MT systems in politically-sensitive cultural/national contexts.

Additionally, these four examples show discrepancies in Meta's response to MT mistakes, depending on who is being affected. For instance, in response to the distorted translations involving prominent world leaders, namely the King of Thailand (Maha Vajiralongkorn) and the President of China (Xi Jinping), Meta took significant action, including apologising, deactivating the auto-translate feature, and offering detailed explanations for the errors. This suggests that Meta gave these incidents more attention due to their high-profile nature even though the impact on Palestinian users, who were not public figures and faced similar (if not worse) consequences, was more severe. Meta's less transparent and uneven responses in these cases point to an ethical double-standard

that is in contrast with Meta's stated Community Standards and content policies that should be applicable equally to all users anywhere (see Meta, n.d.e). These inconsistencies underscore a broader concern that the pursuit of automation and user engagement is outpacing the development of any robust mechanisms for ethical scrutiny and accountability. In the case of MT, ethical responsibility demands more than simply fixing technical issues; it also requires safeguards that ensure fair and equal treatment for all users, regardless of nationality, religion, or political standing. Without such principles in practice, the promise of inclusivity on global platforms remains unfulfilled.

6. Ethical Considerations

EX1 and EX2 raise serious concerns about algorithmic bias and the ethical implications of automated translation systems. Whether the error arose from biased training data, flawed algorithms, or lack of human oversight, the result was the amplification of a harmful stereotype. Given the political sensitivities surrounding the situation in Palestine/Israel, MT tools can inadvertently reinforce existing biases and stereotypes. This reflects a broader concern that automated systems, without proper checks, can reflect the prejudices of the societies that design them. Meta's lack of transparency about how the error occurred fails to address these concerns. Commenting on the Instagram incident of EX1, critics including Palestinian advocates argue that it highlights both digital biases in automated tools and broader issues of content censorship on platforms such as Facebook and Instagram (Paul, 2023). Many Palestinians claim their content has been shadow-banned (shadow-banning is a type of censorship practiced by online platforms where a user's content is hidden or ranked lower, reducing its visibility to others, without the user's knowledge (Suzor et al., 2019), particularly in the context of the ongoing war on Gaza (Paul, 2023). This necessitates that we ask a critical ethical question about how companies such as Meta ensure that their systems are fair, accurate, and non-discriminatory, particularly for individuals who depend on these systems to access information and may be linguistically marginalised, especially in politically sensitive or volatile contexts. These users are more vulnerable to the disproportionate impact of biases in automated systems. As mentioned before, according to Canfora and Ottmann (2020), liability for harm caused by MT mistranslations is still an unresolved issue because machines are not legally liable, leaving users and providers in an ambiguous legal position.

Putting forward ethical guidelines to ensure that MT systems are used responsibly is a necessity, with appropriate safeguards in place to protect vulnerable users from harmful consequences and actions. The way forward could be to explicitly label machine-translated content with a warning, written in plain language, indicating that it may contain errors. This would encourage users to engage with the content more cautiously, reducing the possibility of taking harmful or incorrect actions based on faulty translations. Additionally, users could be given the option to report potentially harmful translations, with a quick review and correction by a human team who can monitor, assess, and resolve translation-related issues. Currently, Facebook users can only rate the quality of a translated post by clicking a star (Facebook, n.d.), offering limited feedback on translation accuracy or appropriateness. Conducting regular evaluation to identify bias and inaccuracies in MT systems is essential, particularly for underrepresented languages with sensitive political or social contexts. However, the feasibility of

implementing these measures remains uncertain given the recent decision by Meta to put an end to its fact-checking program, leaving content moderation to the community without expert oversight or critical evaluation.

7. Corporate Responsibility

By treating the incidents as isolated technical failures, Meta avoids addressing the deeper accountability and ethical challenges that its systems pose, particularly in politically and socially sensitive contexts. These errors can be seen as indicators of underlying structural flaws in the design and deployment of their MT systems. EX1 and EX2 reflect how their MT systems might have been trained on corpora that over-associate Arabic with violence. Instagram's insertion of "terrorist" shows how linguistic bias can be turned into algorithmic violence, dehumanising a whole group of people. The anger and frustration voiced because of the mistranslations demonstrate how such errors can undermine users' confidence in MT systems and their providers. A former Facebook employee with insight into internal discussions regarding Meta's censorship of Palestinian content said the incident (EX1) deeply disturbed many, noting that it can no longer be excused as a technical problem when it spreads misinformation and dehumanises Palestinians (Paul, 2023). While Meta denied intentional censorship, similar claims were made during the major escalation of violence in Gaza in May 2021, when users reported reduced reach, and even removal, of pro-Palestinian posts (Paul, 2021). This incident led to a letter signed by over 200 employees, prompting an independent review commissioned by Meta, which found the company had indeed censored pro-Palestinian content and violated users' rights (Paul, 2023). Furthermore, it has also been reported that Meta applies content moderation unfairly, with Arabic-language posts about Palestine being removed more often than Hebrew content about Israel (Paul, 2024). A recent report by Human Rights Watch (2023), *Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook*, criticised Meta for censoring posts related to Palestinian human rights, despite promises to protect free expression and access to information. Meta has also been criticised for making repeated promises to address these issues, but have often failed to result in meaningful action (Human Rights Watch, 2023). With calls for more clarity on the company's moderation policies, and based on the examples discussed, it can be said that the role of social media giants in shaping public discourse appears to extend beyond moderating original content to also influencing how content is presented through its MT systems.

As a technology subject to ongoing development, MT is currently at the "Peak of Inflated Expectations" stage, which can lead to "a certain blindness regarding the risks" associated with the use of machine-generated content (see discussion in Canfora and Ottmann, 2020, p. 58). Overestimation of MT capabilities can cause users to overlook the need for human supervision and quality assurance measures. Providers of these services need to be transparent regarding their limitations and capacity to ensure that users are aware and can make informed decisions.

As shown in Examples 1-4, MT can be unreliable, especially for languages that are less documented. For example, the English language, due to its history, is overrepresented in datasets used to train AI, whereas languages such as Thai and Burmese lack sufficient representation, making MT less reliable. This is because MT systems are often trained on large datasets that are biased towards dominant languages. Furthermore,

MT is not always effective at understanding slang or cultural context, which can alter the meaning of the source text (Monzó-Nebot & Wallace, 2024). The MT error that led to the unjust arrest of the Palestinian worker shows the extent of the problem of using MT without human supervision. Despite all of these concerns and legitimate debate, tech companies continue to develop and deploy free-to-use MT tools for various purposes. The key question here is what are the ethical implications of making unreliable MT tools available in contexts where they can potentially cause harm? The previous accounts of MT errors with harmful impact highlight that tech companies who provide such services need to:

- Show an ethical obligation to develop and deploy technology that protects vulnerable populations and respects basic human rights.
- Ensure human oversight to reduce the risks of relying on MT systems.
- Carry out frequent assessments of training datasets to identify and mitigate harmful linguistic patterns and underrepresented varieties.
- Use diverse data sources by intentionally including non-dominant language varieties and marginalised community texts.

Regulatory frameworks such as the Digital Services Act (DSA), the UK Online Safety Act (OSA), Ireland's Online Safety and Media Regulation Act, and the EU Code of Practice on Disinformation should be expanded to include content distributed via MT technology. These frameworks prioritise algorithmic transparency, accountability, content moderation and risk mitigation in digital spaces, principles that are essential for promoting responsible AI development, protecting language rights in the digital age and ensuring equal access to information. They can be used to strengthen protection for marginalised communities (socially, politically and linguistically) and vulnerable users, while also increasing corporate accountability. For linguistically marginalised users, MT errors and biased moderation can significantly limit their ability to engage, communicate, and access reliable information. By requiring platforms to carry out impact assessments, improve algorithmic oversight, and establish accessible mechanisms for users to report harms, these regulations can help ensure that all users are treated fairly and protected from digital harm. Furthermore, including specific provisions within the DSA and other similar regulatory frameworks to address linguistic diversity such as requiring evaluation of MT performance by language or mandating human review for high-risk content would close a critical gap. Without such measures, non-dominant language speakers will remain vulnerable to systemic mistranslation and, as a result, misinformation.

8. Concluding Remarks

The four real-life case studies of mistranslations presented have shown that MT can be a double-edged sword, offering convenience but also introducing significant risks, especially if the output is not properly reviewed or contextualised. They have revealed a serious issue with MT when dealing with user-generated content and low-resource languages, where the potential for misinterpretation and misuse is considerable. With its widespread reach and influence, misinformation generated by MT systems on social media has the potential to shape public perceptions, reinforce biases, and lead to harmful action, thus posing substantial risks to the users. Therefore, the responsibility

for ethical AI deployment is a shared one, involving not only developers but also those deploying or disseminating these technologies, e.g., social media providers.

The legal, political, and social consequences associated with these errors highlight the need to address the inherent risks and biases posed by machine-translated content. What we need is more robust, transparent, and ethical approaches to developing, deploying and evaluating MT systems. Social media platforms that also provide MT services should invest in enhancing the quality and context-awareness of their translation algorithms through high-quality training data, neutral data, in multiple languages. It would also be helpful to explicitly label machine-translated content with a warning, written in plain language, indicating that it may contain errors. This might help users better understand the potential downsides of these tools and encourage them to engage with the content carefully, reducing the likelihood of taking harmful or incorrect actions based on faulty translations. In sensitive contexts, they should also provide users with options to report translation errors or offer more human oversight for translations. Human supervision and engagement of end-users must be central to the deployment of AI technologies, particularly when dealing with vulnerable populations to reduce the risks of misrepresentation and harmful stereotypes. This is vital to prevent reinforcing inequalities and power imbalances, such as those between authoritative and vulnerable users, or between more dominant languages versus marginalised ones (see Bender et al., 2021 and Nee et al., 2021 for a discussion on power imbalance in language technologies and linguistic justice). Legal frameworks need to address these issues by promoting responsible AI development and corporate accountability to protect language rights and ensure fair access to information for all users in this digital age.

Future research should critically explore the provision of free-to-use translation services by tech companies and service providers in the context of national and international legal frameworks, aiming to improve access to accurate, multilingual information and promote linguistic justice. This is especially relevant for users in countries with limited access to, or influence over, content created about them, which may still be used to train these systems. Such efforts ensure that users are not just passive recipients of these systems, but active participants in shaping their use and development. Further research may also examine user attitudes towards the reliability of automatic translation tools and platform credibility. This may involve identifying instances where original content was significantly changed in terms of meaning due to mistranslation and interviewing these users to gain insights into their awareness of the errors, their perceptions of the impact, and their trust in the platform's translation tools. This approach can build upon the current study by extending the analysis from the technical and ethical dimensions of mistranslation to user-centred perspectives, exploring how these translation failures affect individual users' sense of agency, identity, and communication.

Acknowledgment

The research conducted in this publication was funded by the Irish Research Council under grant number (GOIPD/2022/341).

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election.

Journal of economic perspectives, *31*(2), 211–236.

Al Sharou, K., Li, Z., & Specia, L. (2021). Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 53–62). Retrieved from https://aclanthology.org/2021.ranlp-1.7/

Al Sharou, K., & Moorkens, J. (2024). Transitude: Machine Translation on Social Media: MT as a potential tool for opinion (mis) formation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* (Vol. 2, pp. 2–3).

Al Sharou, K., & Specia, L. (2022). A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd annual conference of the European Association for Machine Translation* (pp. 171–180).

Baker, S. A. (2022). Alt. Health Influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. *European Journal of Cultural Studies*, *25*(1), 3–24.

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).

Booth, R. (2025). Ditching of Facebook factcheckers a 'major step back' for public discourse, critics say. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2025/jan/07/ditching-facebook-factcheckers-major-step-back-public-discourse

Calzada-Pérez, M. (2014). *Apropos of ideology: translation studies on ideology-ideologies in translation studies*. Routledge.

Canfora, C., & Ottmann, A. (2020). Risks in neural machine translation. *Translation Spaces*, *9*(1), 58–77.

Clark, M. (2022). Google Translate Suggested 'Blow Up The Car' When Arabic Word 'Plan' Entered. *Newsweek*. Retrieved 2024-12-11, from https://www.newsweek.com/google-translate-suggests-blow-car-when-arabic-word-plan-entered-1726537

Cole, S. (2023). Instagram 'Sincerely Apologizes' For Inserting 'Terrorist' Into Palestinian Bio Translations. *404 media*. Retrieved from https://www.404media.co/instagram-palestinian-arabic-bio-translation/

Comito, C. (2023). The role of social media in the battle against COVID-19. In *Artificial Intelligence in Healthcare and COVID-19* (pp. 105–124). Academic Press.

Comito, C. (2024). Polarization and Misinformation: Anticipating Early Signs of Potential Fake News on Social Media. In *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA* (pp. 1–8).

Corbin, C. M. (2017). Terrorists are always Muslim but never white: At the intersection of critical race theory and propaganda. *Fordham L. Rev.*, *86*(455).

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., … Team, N. (2022, August). *No language left behind: Scaling human-centered machine translation*. arXiv. Retrieved from http://arxiv.org/abs/2207.04672 (arXiv:2207.04672 [cs])

Department for Science, Innovation & Technology. (2025). *Online Safety Act: explainer. Updated 24 April*. Retrieved 2025-05-12, from https://www.gov.uk/government/

publications/online-safety-act-explainer/online-safety-act-explainer

Department of Culture, Communications & Sport. (2022). *Online Safety and Media Regulation Act.* Retrieved 2025-12-12, from https://www.gov.ie/en/publication/d8e4c-online-safety-and-media-regulation-bill/

European Commission. (n.d.). *The 2022 Code of Practice on Disinformation.* Retrieved 2025-05-11, from https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

European Union. (2022). *Digital Services Act (DSA) - Regulation (EU) 2022/2065.* Retrieved 2025-04-12, from https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

Facebook. (n.d.). *Translate Feed.* Retrieved 2025-05-11, from https://www.facebook.com/help/509936952489634

Fan, L., White, M., Sharma, E., Su, R., Choubey, P. K., Huang, R., & Wang, L. (2019). *In plain sight: Media bias through the lens of factual reporting.* arXiv. Retrieved from http://arxiv.org/abs/1909.02670 (arXiv:1909.02670 [cs]) doi: 10.48550/arXiv.1909.02670

Farhan, A. K. (2017). *Ideological manipulation in the translation of political discourse: a study of presidential speeches after the Arab Spring based on corpora and critical discourse analysis* (Doctoral dissertation). University of Surrey.

Ghosh, S., & Caliskan, A. (2023). Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society,* (pp. 901–912).

Hern, A. (2017). Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*. Retrieved 2025-01-03, from https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Human Rights Watch. (2023). Meta's Broken Promises Systemic Censorship of Palestine Content on Instagram and Facebook. *HRW*. Retrieved from https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

Instagram. (n.d.). *How Instagram Feed Works.* Retrieved 2025-01-17, from https://help.instagram.com/512686498916530/?helpref=related_articles

Ireton, C., & Posetti, J. (2018). Journalism, 'Fake News' & Disinformation: Handbook for Journalism Education and Training. *UNESCO Publishing*. Retrieved 2025-01-14, from https://webarchive.unesco.org/web/20230926213448/https://en.unesco.org/fightfakenews

Kaplan, J. (2025). More Speech and Fewer Mistakes. *Meta*. Retrieved 2025-05-03, from https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/?utm_source=chatgpt.com

Khurma, M. (2024). A Year of War Since October 7: On Narrative and Dehumanization in Arab and Israeli Societies. *Wilson Center*. Retrieved 2025-06-25, from https://www.wilsoncenter.org/article/year-war-october-7-narrative-and-dehumanization-arab-and-israeli-societies

Lee, K. W., & Qian, M. (2022). Misinformation in Machine Translation: Error Categories and Levels of Recognition Difficulty. In *International Conference on Human-Computer Interaction* (pp. 533–545). Cham.

Marking, M. (2020). Thai Mistranslation Shows Risk of Auto-Translating Social Media Content. *Slator*. Retrieved 2025-05-06, from https://slator.com/thai-mistranslation-shows-risk-of-auto-translating-social-media-content/#:~:text=A%20July%2028%2C%202020%20post,in%20Thai%20on%20Thai%20PBS'

Martin, D. A., Shapiro, J. N., & Nedashkovskaya, M. (2019). Recent trends in online foreign influence efforts. *Journal of Information Warfare*, *18*(3), 15–48.

Martindale, M. J. (2020). Responsible 'Gist'MT Use in the Age of Neural MT. In *Workshop on the Impact of Machine Translation (iMpacT 2020)* (pp. 18–45).

Martindale, M. J. (2024). *When Good MT Goes Bad: Understanding and Mitigating Misleading Machine Translations* (Doctoral dissertation). University of Maryland, College Park.

McMahon, L., & Tidy, J. (2023). Instagram sorry for adding 'terrorist' to some Palestinian user bios. *BBC*. Retrieved 2024-12-01, from https://www.bbc.com/news/technology-67169228

McPherson, p. (2020). Facebook says technical error caused vulgar translation of Chinese leader's name. *Reuters*. Retrieved 2025-01-01, from https://www.reuters.com/article/us-myanmar-facebook/facebook-apologizes-after-vulgar-translation-of-chinese-leaders-name-idUSKBN1ZH0IB/

Meta. (n.d.a). *About.* Retrieved 2025-05-11, from https://www.meta.com/en-gb/about/

Meta. (n.d.b). *No Language Left Behind Driving inclusion through the power of AI translation.* Retrieved 2025-05-11, from https://ai.meta.com/research/no-language-left-behind/

Meta. (n.d.c). *Code of Conduct for Virtual Experiences.* Retrieved 2025-05-11, from https://www.meta.com/gb/legal/quest/code-of-conduct-for-virtual-experiences/

Meta. (n.d.d). *Hateful Conduct.* Retrieved 2025-05-11, from https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/

Meta. (n.d.e). *Community Standards.* Retrieved 2025-05-11, from https://transparency.meta.com/en-gb/policies/community-standards

Meta. (n.d.f). *Misinformation.* Retrieved 2025-05-11, from https://transparency.meta.com/en-gb/policies/community-standards/misinformation/

Monzó-Nebot, E., & Wallace, M. (2024). Gender and ethnolinguistic lawfare: Weaponizing the law. *Just. Journal of Language Rights & Minorities, Revista de Drets Lingüístics i Minories*, *3*(2), 7–116.

Narayanan, S. (2022). Automated misinformation: Mistranslation of news feed using multi-lingual translation systems in Facebook [Abstract]. In *Affinity Workshop: Global South in AI.* Retrieved 2025-05-01, from https://nips.cc/virtual/2022/62672

Nee, J., Macfarlane Smith, G., Sheares, A., & Rustagi, I. (2021). October. Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–9).

Nurminen, M., & Koponen, M. (2020). Machine translation and fair access to information. *Translation Spaces*, *9*(1), 150–169.

Osama, Y. (2025). Deconstructing Narratives: An Analysis of Dehumanization Techniques in the US Media Representation of Palestinians at the Onset of the War on Gaza. *The Arab Journal of Media and Communication Research (AJMCR)*(48), 143–185.

Paul, K. (2021). Facebook under fire as human rights groups claim 'censorship' of pro-Palestine posts. *The Guardian*. Retrieved 2025-01-10, from https://www.theguardian.com/media/2021/may/26/pro-palestine-censorship-facebook-instagram

Paul, K. (2023). Instagram users accuse platform of censoring posts supporting Palestine. *The Guardian*. Retrieved 2025-01-10, from https://www.theguardian.com/media/2021/may/26/pro-palestine-censorship-facebook-instagram

Paul, K. (2024). Meta struggles with moderation in Hebrew, according to ex-employee and internal documents. *The Guardian*. Retrieved 2025-05-11, from https://www.theguardian.com/technology/article/2024/aug/15/meta-content-moderation-hebrew

Pedace, L. (2021). Misinformation and Manipulation on Social Media: User-based and Network-based view. *iSCHANNEL*, *16*(1).

Petrosyan, A. (2025). Number of internet and social media users worldwide as of February 2025 (in billions). *Statista*. Retrieved from https://www.statista.com/statistics/617136/digital-population-worldwide/

Riggs, A. (2019a). On France, terrorism and the English press: Examining the impact of style in the news. In C. Declercq, J. Munday, & M. F. Federici (Eds.), *Intercultural crisis communication: Translation, interpreting and languages in local crises* (pp. 193–214). Bloomsbury Publishing.

Riggs, A. (2019b). *Stylistic deceptions in online news. Journalistic style and the translation of culture* (Bloomsbury ed.).

Saaida, M. (2023). The Role of Social Media in Shaping Political Discourse and Propaganda. *Science for all Publication*, *3*(2), 1–8.

San, M. E., Usanavasin, S., Thu, Y. K., & Okumura, M. (2024). A Study for Enhancing Low-resource Thai-Myanmar-English Neural Machine Translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(4), 1–24.

Sanna, G. A., & Lagnado, D. (2025). Belief updating in the face of misinformation: The role of source reliability. *Cognition*, *258*, 106090.

Serrano, J. (2020). *Facebook Apologizes for Translating Chinese President's Name as 'Mr Shithole'.* Retrieved 2025-01-01, from https://gizmodo.com/facebook-apologizes-for-translating-chinese-president-s-1841095962

Slim, A., & Melouah, A. (2024). Low resource Arabic dialects transformer neural machine translation improvement through incremental transfer of shared linguistic features. *Arabian Journal for Science and Engineering*, 1–17.

Soponvijit, K. (2022). *Lost in translation: Facebook's royal translation error.* Retrieved 2025-01-11, from https://tankytech.net/2021/04/05/facebooks-royal-translation-error

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*(18).

Taylor, J. (2023). Instagram apologises for adding 'terrorist' to some Palestinian user profiles. *The Guardian*. Retrieved 2024-12-16, from https://www.theguardian.com/technology/2023/oct/20/instagram-palestinian-user-profile-bios-terrorist-added-translation-meta-apology

The BBC. (2017). No Arabic-speaking police officer was consulted before the man was

arrested on suspicion of incitement, online:. *BBC*. Retrieved 2025-01-19, from https://www.bbc.com/news/world-middle-east-41714152

The BBC. (2020). Facebook blames 'technical issue' for offensive Xi Jinping translation. *BBC*. Retrieved 2025-01-11, from https://www.bbc.com/news/world-asia-51166339

Tsimitakis, M. (2025). *EU Demands Comprehensive Risk Assessment from Meta Before Fact-Checking Elimination.* Retrieved 2025-05-13, from https://creativesunite.eu/article/eu-demands-comprehensive-risk-assessment-from-meta-before-fact-checking-elimination

Tzoneva, D. (2023). Fixing a Low-Resource Language's Quality Issues — Burmese. *Pulse of Asia*. Retrieved 2025-01-11, from https://www.1stopasia.com/blog/low-resource-language-issues-burmese/

UN. (2022). UN: ARTICLE 19 welcomes report on the right of access to information. *Article 19*. Retrieved 2025-05-01, from https://www.article19.org/resources/un-article-19-welcomes-report-on-the-right-of-access-to-information/

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021). *Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation.* arXiv. Retrieved 2025-10-25, from http://arxiv.org/abs/2102.00287 (arXiv:2102.00287 [cs]) doi: 10.48550/arXiv.2102.00287

Vincent, J. (2022). *Meta open sources early-stage AI translation tool that works across 200 languages.* Retrieved 2025-01-11, from https://www.theverge.com/2022/7/6/23194241/meta-facebook-ai-universal-translation-project-no-language-left-behind-open-source-model

Wang, B., & Feng, D. (2018). A corpus-based study of stance-taking as seen from critical points in interpreted political discourse. *Perspectives*, *26*(2), 246–260.

Warner, A. (2022). *Following outcry, Google Translate removes offensive example phrase for Arabic entry.* Retrieved 2025-04-11, from https://multilingual.com/google-translate-arabic-bias/