Beyond General Purpose Machine Translation: The Need for Context-specific Empirical Research to Design for Appropriate User Trust

WESLEY HANWEN DENG, Carnegie Mellon University, USA
NIKITA MEHANDRU, University of California, Berkeley, USA
SAMANTHA ROBERTSON, University of California, Berkeley, USA
NILOUFAR SALEHI, University of California, Berkeley, USA

Machine Translation (MT) has the potential to help people overcome language barriers and is widely used in high-stakes scenarios, such as in hospitals. However, in order to use MT reliably and safely, users need to understand when to trust MT outputs and how to assess the quality of often imperfect translation results. In this paper, we discuss research directions to support users to calibrate trust in MT systems. We share findings from an empirical study in which we conducted semi-structured interviews with 20 clinicians to understand how they communicate with patients across language barriers, and if and how they use MT systems. Based on our findings, we advocate for empirical research on how MT systems are used in practice as an important first step to address the challenges in building appropriate trust between users and MT tools.

ACM Reference Format:
Wesley Hanwen Deng, Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Beyond General Purpose Machine Translation: The Need for Context-specific Empirical Research to Design for Appropriate User Trust. 1, 1 (April 2022), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine Translation (MT) is widely available as stand-alone translation tools [54, 60] and embedded in larger user-facing systems [8, 9]. Recent advances in neural machine translation have accelerated the progress of MT design and development [3, 60]. Despite the rapid technological improvement, however, MT models can still produce translations that are imperfect or even wrong, causing frustrations or harm to individuals when people over-rely on the outputs [62, 64]. For instance, in 2017, the MT system embedded in Facebook translated a post from a Palestinian man saying "good morning" in Arabic to "attack them," and the poster was arrested by Israeli police [5]. As MT is widely used in high-stakes contexts such as healthcare [27], immigration [50], and law [58], it is important to develop ways to support end-users to build appropriate trust in MT systems [42]. Calibrating trust and reliance could enable more effective human-MT collaboration. At the same time, MT systems should encourage users to fully understand the limits of translation technology to avoid harm caused by over-trust.

Authors' addresses: Wesley Hanwen Deng, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, hanwend@cs.cmu.edu; Nikita Mehandru, University of California, Berkeley, Berkeley, California, USA, nmehandru@ischool.berkeley.edu; Samantha Robertson, University of California, Berkeley, Berkeley, California, USA, samantha_robertson@berkeley.edu; Niloufar Salehi, University of California, Berkeley, Berkeley, California, USA, nsalehi@berkeley.edu.

Manuscript submitted to ACM

In this paper, we first discuss five challenges in calibrating trust between users and MT systems. In particular, users lack intuition and expertise to guide their judgments of translation quality, since they usually have no or limited proficiency in one of either the source or the target language. In addition, populations that might benefit from trustworthy MT systems the most could be hard to reach and engage equitably in research. Adding an additional layer of complexity, it is difficult to communicate translation quality to end-users, since translation is nuanced and context-dependent.

We then present findings from an empirical study in which we conducted semi-structured interviews with 20 clinicians to understand how MT was used in their daily practices and their needs for trust in medical MT systems. We observed that MT served as a low-cost and efficient way of facilitating conversations between clinicians and patients in under-resourced hospitals, but was usually viewed as a last-resort option due to the lack of trust, especially in high-stakes medical settings where patients' consent was required. We also identified techniques clinicians had developed to navigate and calibrate trust in communications with patients that were mediated by human interpreters, which offer some paths forward on designing future trustworthy MT tools. In addition, while current commercial MT systems are mostly general-purpose, clinicians were vocal about having MT systems that are specifically designed to prioritize accurate translation in medical settings and evaluated by a hospital board or medical society.

To conclude, we stress the need for conducting empirical studies to understand how people navigate language barriers and whether and how MT systems are being used in specific domains and contexts. This understanding is an important but currently neglected first step to address the challenges in building appropriate trust between users and MT tools.

2 FIVE CHALLENGES IN BUILDING TRUSTWORTHY MACHINE TRANSLATION

Prior research has surfaced several reasons why developing trust in AI systems, and large language models in particular, is hard. For example, model evaluations based on metrics like the BLEU score [37] is disconnected from real-world application performance that is related to how users perceive the outcome qualities [39]. This gap between evaluation and performance is exacerbated by design asymmetries between developers and users in the current AI product lifecycle. In particular, large language models are usually constructed by practitioners with access to substantial computational resources [4], users whose lives are impacted by these models are being left out in the design and development [44]. Current MT systems based on neural machine translation share this challenge when it comes to calibrating trust between MT systems and users.

In this section, we further detail five distinct challenges that exist in building trustworthy MT systems, based on prior literature from both the MT and HCI communities and our own experience conducting empirical studies on MT systems [31, 42, 43].

2.1 Users lack intuition and expertise to guide their judgments of translation quality

A central challenge in designing trustworthy machine translation systems for everyday users is that these users are usually using MT because they are not fluent in one of either the source or the target language. This means that they have **limited ability to assess the quality of the translations that an MT system has produced and determine whether to rely on them or not.** Prior research studying trust in ML systems and developing interventions to appropriately calibrate trust has largely focused on systems where the user either has expertise in the decision domain and is being assisted by ML, or where some general knowledge or intuition applies [51]. MT poses unique challenges because users often cannot comprehend the output of the system, let alone how it was produced or whether it is correct. To circumvent this challenge, users and researchers have relied heavily on back-translation (translating an output back to the source language) [32, 33, 46, 63]. While this offers some insight into the translation in a language the user can understand, it remains unclear how reliable this strategy is in practice, or in what specific cases it helps or fails to support users in calibrating their trust in MT [33, 67]. For example, Zouhar et al. found that showing users the backtranslation increased their trust in an MT model, but did not help them to identify or improve poor quality translations [67]. This is consistent with evidence in MT and in other ML domains that showing additional interpretability or explainability information can increase users' trust in a model simply because the additional information is there, even if that information indicates that the model may be incorrect [20, 33].

Researchers have also designed and evaluated interfaces that provide additional information other than the back-translation, such as two different translations [15, 61], highlights of key words [14, 26], emotional and cultural context [26], and numerical indications of translation quality [34]. Lab studies suggest that additional information can improve message recipients' perceived understanding without increasing the mental workload [14, 15, 26, 61], but it is not clear from these studies whether users' *perceived* understanding aligns with their actual understanding of the intended meaning [43]. In other words, we do not know whether these interventions help users appropriately calibrate their trust in MT, or whether it promotes trust indeterminately even when translations are incorrect.

2.2 Translation quality is nuanced and context-dependent

To design systems that help a user decide whether to trust a translation or not, researchers need an operational definition of what makes a translation trustworthy. However, translation quality is complex, nuanced, and context-dependent [49]. **In many cases, there are many acceptable translations of a single utterance, while in others there is no translation that truly captures the meaning of a word or phrase in the source language.** Whether a user should rely on a particular translation depends heavily on the context of use. For instance, the same translation error may be humorous in one context, but rude in another. In some circumstances, accurate translation of specific terminology may be crucial, while in others users may prioritize getting the gist of an utterance or enough meaning to move on with an interaction. In each case, the design requirements for trustworthy MT may look very different and it remains unclear when, how, and what kind of quality measures might be intelligible and actionable to users in different situations.

2.3 Even rare and subtle errors can be extremely detrimental to users

Counter-intuitively, another challenge for designing for trust in MT is the relatively high performance of consumer MT systems. Neural MT models have improved substantially in recent years, especially for high resource languages. While we would certainly expect the user experience to improve with higher quality translations, this poses several challenges for engendering appropriate trust. First, as users interact with higher quality translations, their trust in the underlying model is likely to improve. Neural MT models are especially good at producing fluent translations, even if they do not adequately convey the meaning of the source text. Research has suggested that interacting with fluent but inadequate translations does not negatively impact trust, possibly leading users to overtrust a model that produces very fluent output [30]. Second, reports of excellent performance may lead researchers and developers to deprioritize the development of more trustworthy MT systems, believing that there is little need to warn users of errors if those errors are rare and insignificant [4].

However, **particularly in high-stakes settings, rare errors can have drastic consequences**. For instance, Khoong et al. used Google Translate to translate emergency department discharge instructions into Spanish and Chinese [22]. While the authors found that most translations were acceptable quality, a small number contained potentially life-threatening mistakes, for example, when "hold the kidney medicine until you have a chance to speak with your kidney doctor," was translated to "keep taking kidney medicine until you talk to your kidney doctor" in Chinese. Further, MT performance in high-resource languages has advanced far more quickly than performance in lower resource languages. The need to support reliable use of MT in low-resource languages is made even more urgent by this performance disparity, as users may not be aware of these differences and may not adjust their level of trust in the system when switching from a higher-resource to a lower-resource language pair.

2.4 It is challenging to conduct ecologically valid empirical studies of MT

Recent studies on trustworthy AI emphasize the importance to design and conduct sound empirical research on how AI systems are being used (and misused) in real-world scenarios in order to capture a well-rounded and comprehensive understanding of the trust construct [16]. However, **people who might benefit the most from trustworthy MT are also those who tend to be hard to reach or collaborate with**. For example, when Liebling et al. tried to understand how migrants in India used MT, female participants were reluctant to be interviewed without their husbands present or rejected the interview without permission from their husbands [25], biasing their participant sample towards male participants. There is also an epistemological burden for conducting research studies with community members, especially to marginalized social groups or people working in high-stakes jobs, as they are often short on time and resources [18, 38]. This challenge was reflected in one of the authors' study designs in which we attempted to conduct an ethnographic study with COVID testing volunteers who use MT to help communicate with local immigrants. We ultimately decided to change the study design realizing that our study would impose burdens on already over-worked volunteers and their hectic working schedule.

2.5 High dimensional design space

Currently, the most popular machine translation tools, like Google Translate or Microsoft Translator, are general purpose tools that serve a wide range of users involved in a wide range of tasks across a wide range of languages. However, as each of the above challenges highlighted, **designing for *trustworthy* MT depends heavily on who is using an MT system, and for what**. MT performance varies widely across language pairs, and across dialects of a language [10]. Some users have proficiency in source and target language, and are using MT for only a word or phrase they do not know how to translate, while others completely lack knowledge of one of the languages. People use MT in a huge diversity of contexts, from low-stakes settings like reading foreign news or getting directions while traveling, to high-stakes settings like healthcare, content moderation, and policing [5, 22, 53, 57]. Each of these contexts will call for different kinds of support for users to judge when it is ethical and appropriate to rely on MT. Thus, we have begun by narrowing in to understand how clinicians use MT in their practice, whether they are able to do so reliably, and how we might be able to help them calibrate their trust in translation technologies so that they can provide language support more safely.

3 CASE STUDY: USING MACHINE TRANSLATION IN MEDICAL SETTINGS

While professional medical interpreters remain the gold standard for facilitating cross-lingual communication in medical practice, finding skilled translators in a timely manner can be difficult and time-consuming. Clinicians have resorted to machine translation tools when short for time and when operating in low-stakes medical settings [35, 68]. Although there are safety and reliable concerns around machine translation (MT) system deployment in medical settings, MT can be a low-cost and efficient way of communicating with patients in under-resourced hospitals and clinics where there are limited or no medical interpreters. To minimize risk in this setting, there is a great need to build more trustworthy MT system to facilitate the conversation between clinicians and patients when they don't speak the same language.

To this end, we conducted an interview study with 20 clinicians, including physicians, surgeons, nurses, and midwives, in the U.S., as an initial step to understand how clinicians are currently interacting with machine translation systems. Our study included clinicians from across seven medical specialities: cardiology, orthopedic surgery, nephrology, family medicine, obstetrics and gynecology, trauma surgery, and emergency medicine. We found that medical literacy rates among patients, cultural barriers, and time and resource constraints greatly impact patient-clinician communication. Building trustworthy systems that ultimately account for reliability and transparency can improve the quality of care of patients in low-resource settings, and can complement existing language translation services in high-resource settings [21]. Our findings have important implications for how we might address the challenges above in the design of trustworthy machine translation systems for clinician use.

3.1 Cross-lingual communication Medical Settings

Generally, clinicians in our sample turned to professional medical interpreters when they needed to communicate with a patient with whom they did not share a language. However, challenges like time constraints and limited interpreter availability led some to occasionally resort to machine translation. In this section, we outline how and when clinicians used MT in their practice. In the next section, we explore how they tried to evaluate and calibrate their trust in both human interpreters and MT and identify unmet needs and opportunities for designing trustworthy translation support for medical settings.

Some clinicians only used machine translation tools when they had baseline familiarity, but not necessarily medical fluency, with the target language. In this case, they used their language ability to evalute the accuracy of the translation.

> *I've only really used Google Translate in clinic, in a situation where I'm counseling a patient, and maybe it's on an issue that I just haven't spoken to a patient in Spanish about recently. And so, I just want to make sure, am I using the correct word? I mean, essentially, if I plug it in and I'm looking in Spanish, I can tell whether the translation is correct or not. Whereas in other languages, I can't, which is why I don't use Google Translate for other languages.* (P4, Obstetric and Gynecological Surgeon)

A few other clinicians noted that some circumstances, including being short on time and having no access to a language translation line, necessitated the use of Google Translate. In such instances, MT was viewed as a last-resort option.

> *Usually, when I'd use it (Google Translate) is out of sheer desperation. So often it was more rare languages where there was no interpreter, we'd be on hold for 15 minutes, realized we were probably not going to get someone at 6:00 AM, while bedside rounding, and just used Google Translate to do the best we could to try to communicate in that setting.* (P19, Family Medicine Physician)

Aware that both medical interpreters and MT systems may sometimes cause miscommunication with patients, clinicians had developed techniques to improve communication, and evaluate translation quality and patient understanding, including: **a) rephrasing medical terms, b) using back translation techniques, c) relying on nonverbal patient cues, and d) testing patient understanding.** When interacting with patients, clinicians often used simple words in their communication. Clinicians frequently employed back translation methods in which they would ask an interpreter to repeat back what had been translated to the patient, often measuring the time the interpreter spent conversing with the patient versus translating the information back. Nonverbal cues from the patient including gestures and facial expressions were often used to gauge the quality of the interpreter. Lastly, teach back methods are a common practice in most training programs (e.g. medical schools), where clinicians are encouraged to ask the patient to repeat back what they said in terms of the patient plan moving forward. Clinicians frequently employed this technique during cross-lingual patient interactions.

3.2 Need for Improvements in MT Systems Deployed in Medical Settings

Clinicians recognized the limitations of machine translation tools. For example, the citing the tools' inability to pick up context like *"emotional body language cues,"* thus making them only suitable for *"quick [and] simple communication."*

For this reason, clinicians judged the appropriateness of MT on a case-by-case basis, depending on their level of trust in the MT system and the stakes of the interaction. For example, when patient consent was required, clinicians took the time to call a language translation service or a medical interpreter, while in situations they perceived to be lower-stakes, such as taking a patient's medical history, clinicians were more likely to resort to Google Translate to navigate language barriers.

> *I think if it was a more detailed or complex situation, I probably would have called the interpreter. Just again, because legal liability as to whether the patient understood or not.* (P2, Nephrologist)

Clinicians in our study also recognized that general purpose machine translation systems such as Google Translate were not built for use in medical settings. One clinician described:

> *Even something like Dragon Software, which I don't currently use, but I've used in the past for medical dictation. It's not perfect, so if there's something that's geared towards medical language is important, and that it's been checked out by somebody else* (P4, Obstetric and Gynecological Surgeon)

Another clinician noted how translation systems often do not account for dialect differences and tone, including local dialects or slangs, cultural differences reflected through tone or inflection, and sarcasm.

Thus, lack of trust was directly tied to issues of reliability and accountability. Since translation systems offer little feedback on the quality of their translations, clinicians are advised to use these tools cautiously. However, this guidance is generally vague and **it is not clear how effectively clinicians are able to exercise caution when using MT.** Part of this issue is due to the fact that MT tools are not designed specifically for use in healthcare, and evaluations of these systems on medical language suggests a potential for grave harm from translation errors [58]. As such, many were wary of the tool's accuracy in relation to translating medical language and described accountability concerns, noting that when navigating high-stake medical settings such as soliciting patient consent a lack of trust in the MT system prevented them from relying on these tools. As one physician stated:

> *If it was very complex, where I needed patient's consent, probably would've used a translator, because I knew they would not be supported by Google. Nobody would back me up for the Google Translate.* (P2, Neprologist)

The importance of standardizing evaluation methods of machine translation tools within medical settings was a recurrent theme in our study. Training models on medical phrases can build clinician confidence in MT systems. Clinicians also repeatedly commented on the importance of a machine translation tool having been vetted by a medical society to help them build trust in the MT systems. Medical society validation is crucial because it's typically a signal that the tool was tested via a randomized control trial, and that the findings were published in a peer-reviewed medical journal.

> *If it's gone through a vigorous screening process and research and make sure that they're correctly translated and somebody presents to me as a certified medical translation tool, I don't think I would question that.* (P3, Obstetric and Gynecological Surgeon)

In summary, clinicians recognized some of the limitations of MT systems, but lacked concrete guidelines around when to use these tools in medical settings, or feedback about how they could rephrase terms or otherwise adjust how they use the system to improve translation accuracy.

4 TOWARDS TRUSTWORTHY MACHINE TRANSLATION SYSTEMS

Our case study showcased challenges clinicians encountered in their day-to-day practice while using MT and strategies they adopted to try to overcome these challenges. These empirical findings offered new perspectives and opened up conversations on how might we design MT systems for users to build and calibrate trust towards MT, potentially addressing the challenges we mentioned in section 2. To this end, we suggest that an important first step towards building trustworthy MT systems is to **conduct empirical studies to understand how people navigate language barriers and whether and how MT systems are being used in specific domains and contexts**. Within the various domains where MT is used (e.g., medicine, law, social media etc.), user needs vary based on the context, for example, the stakes and formality of the interaction, the communication modality, potential cultural and professional gaps, users' language proficiency, and their familiarity with the technology.

The motivation for this approach is three-fold: *First*, understanding the domain-specific needs and challenges of using MT would help researchers and developers **clarify what user "trust" in MT actually means in each specific context**, which is currently understudied in the MT research community. For example, calibrating a clinicians' trust in MT systems in high-stakes medical contexts is different from the level of trust a tourist may need when using an MT system to help them check into a hotel while traveling in a foreign country. In our case study, clinicians' trust towards MT systems was influenced both by their day-to-day interactions with the system and the institutional endorsements based on expert peer-review systems. Other stakeholders under the same context (e.g., patients) and users in other MT use cases might construct and calibrate their trust in drastically different ways based on how often they use MT, whether the system has been rigorously evaluated, and the importance of accurate communication to the task at hand. Prior research has also found that people's trust towards AI systems is shaped by organizational process [28, 65], culture [17], and power dynamics [52] when cross-functional actors were collectively interacting with an AI system [40]. Contextualizing user "trust" in MT in real-world practice could advance us a step closer to providing nuanced, situated indicators of translation quality (Challenge 2.2), and understanding how to prevent or help users identify and cope with rare but high impact errors (Challenge 2.3)

Second, researchers and developers could **design interfaces and interactions that build upon MT users' current practice and strategies surfaced from the empirical research**. HCI has a long tradition of drawing insight and inspiration from users to design ecologically valid products [11]. Our semi-structured interviews, for example, surfaced that clinicians currently use both back translation techniques and "teach back methods" in their daily communications with patients mediated by human interpreters. Future designs could leverage and incorporate the interactive and reciprocal strategies that clinicians have already developed to improve communication with patients (Challenge 2.1, 2.4). Besides semi-structured interviews, other methods like participatory design [24], diary studies [41], value-sensitive design [13], and speed dating [66] (and the combination of these methods) all have the potential to collect useful data in different formats to help designer future MT systems based on stakeholders' needs (Challenge 2.4).

It is also possible that, through analyzing and making sense of fieldwork data, researchers could surface the latent needs and unspoken desires reflected through users' current practice and interaction patterns, and thus inform future design [59]. For example, there was general acknowledgment among clinicians that MT tools should be given to patients so they can communicate their symptoms to clinicians instead of always having clinicians holding the MT tools. Having patients type their symptoms and medical questions into a tool on their own could yield a more comprehensive medical history. This observation leads us to the design implication that users' *personal agency* while using MT plays an important role in trust-building.

Third, studying MT systems' use in specialized contexts could help researchers prioritize **designing and developing domain-specific, context-aware MT systems.** Current commercial MT systems are mostly general-purpose [60], but trustworthy MT systems can be specifically designed to prioritize accurate translation of the specific language used under the specific context, possibly through domain adaptation methods [12]. Unfortunately, domain-specific model-building and quality evaluation for MT systems are currently understudied. Field data from users could potentially provide useful training datasets and instantiate theoretical frameworks like context-aware learning for neural machine translation [19], narrowing down the design space for MT (Challenge 2.5). MT researchers could also draw from prior HCI research in context-aware ubiquitous computing [45] and mixed-initiative interfaces [2] to design MT tools that are beyond text-based input-output box translation interfaces (Challenge 2.1). Another promising future direction that could complement current progress in neural machine translation is leveraging relatively standardized patterns and topics of conversation to introduce fixed phrase-based translations [36, 47, 55](Challenge 2.5)

In addition to building domain-specific models, it is also crucial to draw from empirical findings to create both general and domain-specific user guidelines [1, 29], training and onboarding materials [6, 7] for MT users, and rigorous evaluation and endorsement [23, 48, 56]. In short, we see value in building domain-specific MT ecosystems.

5 CONCLUSION

In this position paper, we first synthesize challenges in building trustworthy Machine Translation systems. Then, we share a case study in which we conducted semi-structured interviews with 20 clinicians to understand how MT was used in their daily practices and their needs for trust in medical MT systems. As a conclusion, we call for conducting empirical studies to understand how people communicate across languages and how MT systems are being used under specific contexts as an important but often neglected step towards building trustworthy MT systems. We explained how this approach could potentially address the unique challenges in building trustworthy MT, followed by design implications that are generalizable to MT systems for contexts beyond medical settings.

REFERENCES

[1] 2021. People AI Guidebook. (2021). https://pair.withgoogle.com/guidebook/

[2] James E Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2015).

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*. 610–623.

[5] Yotam Berger. 2017. Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'. *Haaretz* (Oct 2017). https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

[6] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213* (2020).

[7] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[8] Facebook Help Center. [n.d.]. *Translate Facebook APP*. https://www.facebook.com/help/364760868183100

[9] Twitter Help Center. [n.d.]. *Tweet Translation*. https://help.twitter.com/en/using-twitter/translate-tweets

[10] Prithwijit Das, Anna Kuznetsova, Meng'ou Zhu, and Ruth Milanaik. 2019. Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English Proficiency Caregivers. *Clinical Pediatrics (Phila)* 58, 2 (Feb 2019), 247–249. https://doi.org/10.1177/0009922818809494

[11] B-N Sanders Elizabeth and Colin T William. 2002. Harnessing people's creativity: Ideation and expression through visual communication. *Focus groups: Supporting effective product development* 137 (2002).

[12] Chelsea Finn, P. Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*.

[13] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[14] Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, Paris, France, 449–458. https://doi.org/10.1145/2470654.2470719

[15] Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, Vancouver, BC, Canada, 852–863. https://doi.org/10.1145/2675133.2675197

[16] Felix Gille, Anna Jobin, and Marcello Ienca. 2020. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intelligence-Based Medicine* 1 (2020), 100001.

[17] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[18] Yen-Chia Hsu, Ting-Hao'Kenneth' Huang, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, and Alessandro Bozzon. 2021. Empowering Local Communities Using Artificial Intelligence. *arXiv preprint arXiv:2110.02007* (2021).

[19] Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715* (2019).

[20] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[21] Elaine C Khoong and Jorge A Rodriguez. 2022. A Research Agenda for Using Machine Translation in Clinical Medicine. *Journal of General Internal Medicine* (2022), 1–3.

[22] Elaine C Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA internal medicine* 179, 4 (2019), 580–582.

[23] Elaine C. Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernández. 2019. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine* 179 (2019), 580–582.

[24] Sarah Kuhn and Michael J. Muller. 1993. Participatory design. *Commun. ACM* 36 (1993), 24–28.

[25] Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376261

[26] Hajin Lim, Dan Cosley, and Susan R. Fussell. 2018. Beyond Translation: Design and Evaluation of an Emotional and Contextual Knowledge Interface for Foreign Language Social Media Posts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3173791

[27] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1, 6 (2019). https://doi.org/10.1016/S2589-7500(19)30123-2

[28] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *arXiv preprint arXiv:2112.05675* (2021).

[29] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[30] Marianna Martindale and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Association for Machine Translation in the Americas, Boston, MA, USA, 13–25. https://www.aclweb.org/anthology/W18-1803

[31] Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, South Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA.

[32] Mai Miyabe and Takashi Yoshino. 2009. Accuracy Evaluation of Sentences Translated to Intermediate Language in Back Translation. In *Proceedings of the 3rd International Universal Communication Symposium* (Tokyo, Japan) *(IUCS '09)*. Association for Computing Machinery, New York, NY, USA, 30–35. https://doi.org/10.1145/1667780.1667787

[33] Mai Miyabe and Takashi Yoshino. 2010. Influence of Detecting Inaccurate Messages in Real-Time Remote Text-Based Communication via Machine Translation. In *Proceedings of the 3rd International Conference on Intercultural Collaboration* (Copenhagen, Denmark) *(ICIC '10)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1841853.1841863

[34] Mai Miyabe and Takashi Yoshino. 2011. Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?. In *Human-Computer Interaction. Interaction Techniques and Environments*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–377.

[35] Tom Moberly. 2018. Doctors choose Google Translate to communicate with patients because of easy access. *BMJ* 362 (2018). https://doi.org/10.1136/bmj.k3974 arXiv:https://www.bmj.com/content/362/bmj.k3974.full.pdf

[36] Anita Panayiotou, Kerry Hwang, Sue Williams, Terence W H Chong, Dina LoGiudice, Betty Haralambous, Xiaoping Lin, Emiliano Zucchi, Monita Mascitti-Meuter, Anita M Y Goh, Emily You, and Frances Batchelor. 2020. The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing* 29, 17-18 (Sep 2020), 3516–3526. https://doi.org/10.1111/jocn.15390

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

[38] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. 2021. Getting Ourselves Together: Data-centered participatory design research & epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[39] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *WMT*.

[40] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *arXiv preprint arXiv:2006.12358* (2020).

[41] John Rieman. 1993. The diary study: a workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 321–326.

[42] Samantha Robertson, Wesley Deng, Timnit Gebru, Margaret Mitchell, Daniel J Liebling, Michal Lahav, Katherine Heller, Mark Díaz, Samy Bengio, and Niloufar Salehi. 2021. Three Directions for the Design of Human-Centered Machine Translation. *First Workshop on Bridging Human–Computer Interaction and Natural Language Processing at EACL 2021* (2021).

[43] Samantha Robertson and Mark Díaz. 2022. Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, South Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA.

[44] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891* (2019).

[45] Bill Schilit, Norman Adams, and Roy Want. 1994. Context-aware computing applications. In *1994 first workshop on mobile computing systems and applications*. IEEE, 85–90.

[46] Tomohiro Shigenobu. 2007. Evaluation and Usability of Back Translation for Intercultural Communication. In *Usability and Internationalization. Global and Local User Interfaces*, Nuray Aykin (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 259–265.

[47] Hervé Spechbach, Ismahene Sonia Halimi Mallem, Johanna Gerlach, Nikolaos Tsourakis, and Pierrette Bouillon. 2017. Comparison of the quality of two speech translators in emergency settings : A case study with standardized Arabic speaking patients with abdominal pain. In *Proceedings of European Congress of Emergency Medicine (EUSEM 2017)*. Athens, Greece. https://archive-ouverte.unige.ch/unige:100812

[48] Hervé Spechbach, Ismahene Sonia Halimi Mallem, Johanna Gerlach, Nikolaos Tsourakis, and Pierrette Bouillon. 2017. Comparison of the quality of two speech translators in emergency settings : A case study with standardized Arabic speaking patients with abdominal pain.

[49] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Sofia, Bulgaria, 79–84. https://www.aclweb.org/anthology/P13-4014

[50] Steve Stecklow. 2018. Why Facebook is losing the war on hate speech in Myanmar. *Reuters* (Aug 2018). https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

[51] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 74, 16 pages. https://doi.org/10.1145/3411764.3445088

[52] Lauren Thornton, Bran Knowles, and Gordon S. Blair. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[53] Yeganeh Torbati. 2019. Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees. *Pro Publica* (September 2019). https://www.propublica.org/article/google-says-google-translate-cant-replace-human-translators-immigration-officials-have-used-it-to-vet-refugees

[54] Google Translate. [n.d.]. *Google Translation.* https://translate.google.com/

[55] Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health Surveill* 5, 1 (Jan 2019). https://doi.org/10.2196/11171

[56] Anne M. Turner, Yong Kyung Choi, Kristin N. Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health and Surveillance* 5 (2019).

[57] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society* 0, 0 (2020), 1–18. https://doi.org/10.1080/1369118X.2020.1776370 arXiv:https://doi.org/10.1080/1369118X.2020.1776370

[58] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society* 24, 11 (2021), 1515–1532.

[59] Colin T. William. 2001. Harnessing People's Creativity: Ideation and Expression through Visual Communication.

[60] Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv* abs/1609.08144 (2016).

[61] Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving Machine Translation by Showing Two Outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3743–3746. https://doi.org/10.1145/2556288.2557171

[62] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009).

[63] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in Establishing Common Ground in Multiparty Groups Using Machine Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/1518701.1518807

[64] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *CSCW '06*.

[65] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[66] John Zimmerman and Jodi Forlizzi. 2017. Speed dating: providing a menu of possible futures. *She Ji: The Journal of Design, Economics, and Innovation* 3, 1 (2017), 30–50.

[67] Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation Feedback Improves User Confidence in MT, Not Quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 151–161. https://doi.org/10.18653/v1/2021.naacl-main.14

[68] Orhan-Sungur Mukadder Özkan-Seyhan Tülay Şentürk, Emre. 2021. Google Translate: Can It Be a Solution for Language Barrier in Neuraxial Anaesthesia? *Turkish journal of anaesthesiology and reanimation* 49, 2 (2021), 181–182. https://doi.org/10.5152/TJAR.2021.101