'I am not a number': On quantification and algorithmic norms in translation

Joss Moorkens, Dublin City University

Numbers and measurements enable transactions and communication in translation in ways that are helpful and indisputably necessary. However, as deployment of quantification and mathematisation has become more complex and opaque, it is important to interrogate the validity of measures and predictions, especially if they are to be used as a basis for action. This article takes a critical look at the various types of quantification and mathematisation used in translation and considers the effects of these on translators working in highly technologized workflows. It introduces the concept of algorithmic norms, whereby translators feel pressured to reverse engineer and conform to the demands of algorithmic management.

A fear regularly expressed in media and literature is of a person being 'reduced to a number', a simplified statistic. However, in recent years as our lives have become digital, our activities are increasingly quantified and our behaviours modelled. This is particularly true in highly technologized contemporary translation production, where words and characters, actions and inactions are digitised, enumerated, quantified, evaluated, and mathematised in order to analyse production steps, automate decisions, and to maximise efficiency. Digitisation is relatively straightforward for the functional mechanics of character encoding and network packet routing, but when followed with quantification, abstraction, and evaluation, can have distorting effects, building an incomplete or biased representation of a text or process. This article is about this potential for distortion through quantification and mathematisation, particularly in the use of algorithms geared to simplify the complexity of systems and processes.

The urge to quantify translation comes not only from business for pricing and commodification, from engineering for retasking, leveraging, and automation, but also from academic research. Studies on machine translation (MT) are usually based on quantitative results, and the growing use of quantitative data in translation studies has led to the use of statistical methods to make sense of those data (Mellinger and Hanson, 2017). The use of quantitative methods is not in itself problematic, but quantification of words and language is an inherently difficult task, as language does not fit comfortably into categories and metrics. Zanettin (2013), for example, has discussed the complex combination of qualitative and quantitative approaches required to make generalisations based on language data. Marais and Meylaerts (2018, p. 2) argue that translation studies has been wedded to a reductionist model, 'decomposing systems into elementary, simple units'. They propose instead that analysis should be 'focused not on parts but on the relationships and connections between parts and between parts and wholes' (Marais and Meylaerts, 2018, p. 9).

Quantification and mathematisation have become almost indispensable in highly technologized translation processes, along with a tendency to focus on decomposed elements rather than the whole. Much of this work takes place on digital platforms, within which machine learning is used to automate project management steps, for MT, and for evaluation (Fırat, Gough and Moorkens, 2024). The following sections look at digitisation and quantification, bringing in some discussion of quantification from the field of philosophy. The notion of algorithmic norms is proposed as an example of the potential mismatch between the needs of end users of translation and the use of algorithms to automate translation workflow steps. The analysis presented draws on the philosophy of science, sociology, and validity theory in order to help answer questions of measurement validity regarding uses and effects. These measures and predictions include translator activity metrics, MT, translation quality evaluation, and translator job allocation. Isolated measures and scores affect business decisions and pricing, prompting decisions that are unlikely to adequately consider sustainability (ecological and social) and may tell an incomplete story. Since these measures have value implications and are used as a basis for action (by automating steps such as employment decisions, for example), we should carefully consider their validity.

Digitisation and quantification

The first step in working with technology is digitisation, the conversion of text and media to a machine-readable series of numbers. As we type within an editing interface, words and characters are digitised according to a defined character encoding standard. If they are to be transmitted, they will go through a number of transformations as described by the Open Systems Interconnection model (International Organization for Standardization, 1994) with sequences of information chopped into data packets and transmission protocol headers appended to guide their paths through cables, switches and routers, to be recombined at another location. In this way the 'production and transmission of data explicitly links the digital universe of physical infrastructure and capacity to the digital world that has materialized through human intervention' (Folaron, 2012, p. 15). Our view at the upper 'Application Layer' of the OSI Model hides the transformation occurring beneath the surface. These transformations do not change our characters and words noticeably, as long as our data conform to standards and expectations. If they do not conform, things quickly become complicated.

Once information has been digitised, it may be more easily quantified (i.e. counted or measured). This might be expected to be a simple exercise, however even something as ostensibly unambiguous and simple as word counting can present difficulties. Words have become the most common basis for cost estimates and payment for translation and localisation work, with number of lines used for agglutinative languages and character counts for many Asian languages (Levitina, 2011). Zydroń (2014, p. 33) highlights how word processing tools count words differently, 'not only between rival products, but also sometimes between different versions of the same product.' Zydroń (2017) provides an example of a complex file with a count of 430 words according to Microsoft Word 2000 and 764 words according to the 2010 version of Word and explains that the measure of numbers of characters from many Asian languages that are counted as a word is similarly inconsistent. These issues with counting can be addressed with standards, assuming that they are universally or fully adopted.

Quantification and Philosophy

The urge to measure predates modern times, as described by Vincent (2022), but became widespread in the post-Enlightenment era, when rationalism and scientific measurement began to be considered as the root of civilisation and insight. The growth in popularity of the philosophical position of positivism, based on a belief only in what can be objectively measured or proven, led to revolutionised industrial production in the early part of the 20th century. This was criticised by mid-century philosophers such as Ellul (1964, p. xxv), who referred to the technologised rationalisation of all parts of our lives as *technique*, which he defined as 'the totality of methods rationally arrived at and having absolute efficiency (for a given stage of development) in every field of human activity'. Heidegger called this focus on measurement, technology and economics 'calculative thinking', which 'computes ever new, ever more promising and at the same time more economical possibilities' (1966, p. 44).

As identified by Ellul, the trend towards quantification moved beyond items that are directly or obviously measurable to concepts that are more abstract, what Galileo called secondary qualities (Berghofer, Goyal and Wiltsche, 2021), in order to give them a comparative value for commodification and exchange or due to an 'identification of the thoroughly mathematized world with Truth' (Horkheimer and Adorno, 1947, p. 18). Horkheimer and Adorno (1947, p. 4) criticised this tendency within a society 'ruled by equivalence', that 'makes dissimilar things comparable by reducing them to abstract quantities.' Duhem (1954, p. 14) predicted that such a process could only entail loss, that 'there cannot be complete parity' as the abstraction 'cannot be the adequate representation of the concrete fact'. The controversy when music was first digitised, with sound waves quantised to ones and zeros, is an example of the computational necessity to 'striate otherwise-smooth… analog details' (Golumbia, 2009, p. 11). Another example could be words or text. Piper (2018, p. 101) finds that computationally modelling a text can give us insights about a single aspect of that text – for example, predictive models using machine learning 'allow us to engage in the process of classification, of what it means to define a group of texts as a coherent entity… according to certain predefined conditions' – but cannot be definitive.

Proposals from Shannon (1948) and Chomsky (1956) to use mathematical notation for language in order to model communication or linguistic structure were criticised for their inability to represent language definitively, but despite this the tendency to mathematise has grown with the popularisation of technology. Husserl coined the term mathematisation for the mathematical representation of what is not directly measurable (Berghofer, Goyal and Wiltsche, 2021). Skovsmose (2020, p. 605) defines mathematisation as 'the formatting of production, decision-making, economic management, means of communication, schemes for surveilling and control, war power, medical techniques, etc., by means of mathematical insight and techniques'. Thus, as Golumbia (2009, p. 14) writes, 'mathematical calculation can be made to stand for propositions that are themselves not mathematical, but must still conform to mathematical rules.'

Mathematisation for prediction

The use of mathematisation for translation prediction has become commonplace since the use of linguistic rules for MT was superseded by the use of data-driven methods. Criticism of this paradigm shift echoed disagreements between rationalists, who believe that certain principles were fundamentally true, and empiricists, who feel that only what can be measured can be true. One review of an early Brown et al. paper on statistical MT complained that 'the crude force of computers is not science' (Way, 2010, p. 181). When neural MT became the leading data-driven MT paradigm in 2016, we moved from what is known as symbolic to subsymbolic artificial intelligence (AI), from using computer representations that are readable to an unexplainable 'black box' system.

Mitchell (2020, p. 24) describes subsymbolic programmes as 'essentially a stack of equations – a thicket of often hard-to-interpret operations on numbers'. Chaitin (2013, p. 33) writes that computing has evolved to a state of 'infinite complexity' that is revolutionary and 'reveals a new world'. The sheer quantity of operations renders the static dynamic. Word embeddings map syntactic relationships between words and phrases, reifying these relations to numerical representations within a vector space. The 20th century criticisms of Duhem and others about the dangers of abstraction feel instinctively true, yet we find that neural MT outperforms other approaches, particularly within narrow domains, to the extent that claims have been made of parity with the quality of human translation (e.g. Hassan *et al.*, 2018) and even of MT outperforming humans (Popel *et al.*, 2020). Using similar processes but at a larger scale, generative tools using large language models have quickly reached similar levels of quality for well-supported languages, with reports of improved capabilities regarding context (Castilho *et al.*, 2023; Hendy *et al.*, 2023). It would thus appear that prediction for translation is inarguably the best path for us to follow, assuming that enough data is available for system training.

There are, however, also some reasons to be less optimistic. The scale of abstraction and mathematisation means that translation processes are opaque. While in general, more data leads to better output, qualitative improvement due to added data appears to degrade, with each similarly-sized addition of data having a lesser effect than the last (Schwartz *et al.*, 2020, p. 56). Data augmentation can also be unpredictable, with improvements or disimprovements from different tranches of data produced without a clear pattern. Crawled data from the web often requires a good deal of curation, with much of it already machine translated, and its addition to training data may not bring about expected improvement. Back-translation of target text data will contain translation errors, but appears on the whole to improve output quality. Moving on from data, multilingual MT seems to improve output quality for poorly-supported languages without a great deal of loss for major languages, and is in favour with research groups from big tech companies (Bapna *et al.*, 2022; NLLB team *et al.*, 2022). The reason for this qualitative improvement is not entirely clear. A contribution from Google in 2017 suggested that the multilingual system created a hidden interlingua, based on the idea of language universals that harks back to the early prognostications of Weaver (1949). MT can very occasionally produce output that is fluent but entirely incorrect or with inexplicably wrong words or negations, known as hallucinations (Guerreiro, Voita and Martins, 2022). There is also no reliable method to remove bias, such as gender or racial bias, from MT output or to consistently produce gender-neutral content (Vanmassenhove, Emmery and Shterionov, 2021). As with Hume’s problem of induction (Henderson, 2022), any translation prediction is necessarily based on previous material, with all of the bias that this entails and no flexibility to update without additional data. With this in mind, translation prediction seems less progressive than we might have believed.

Perhaps this might not be a problem for low-stakes translation, where risk, value, and time is minimal. Unpredictable MT is still not ideal and, as highlighted by Vieira et al. (2021), unrealistic expectations of MT lead to its use in high-stakes situations. Two related solutions to help us to understand what happens within a neural MT system and thus to predict correct translations more accurately involve explainable AI and neurosymbolic approaches. A widely-used explainable machine learning method is to replicate a black-box system in order to explain what happens within the box, an approach that Rudin (2019) says can produce incorrect or misleading results. Instead, she proposes the use of systems that are inherently explainable, adding that there is not necessarily a cost to performance in doing so. She also adds that developers rarely seek the simplest system for a machine learning task, assuming instead that complexity will lead to better results. Her proposed rule, particularly for high-stakes scenarios is that ‘no black box should be deployed when there exists an interpretable model with the same level of performance’ (Rudin 2019, p. 210). Van Harmelen (2022, p. v) believes that developers of symbolic and subsymbolic systems are ‘converging on the view that neither purely data-driven nor purely knowledge-driven systems alone hold the key to further progress in AI’. Marcus (2020) sees this hybrid approach as a first step towards trustworthy and robust AI, thus inferring that current approaches are neither trustworthy nor robust.

Quantification for evaluation

The claim that MT has reached parity with human translation quality by Hassan et al. (2018) is an example of both cherry-picking of translation evaluation approaches and a value judgement with social consequences. Evaluation was carried out using the direct assessment method (Graham *et al.*, 2016), which was previously used solely for comparing systems in competitive MT tasks, via crowdsourcing without consideration of cohesion at the document level. An error-type evaluation included towards the end of the paper did not support the claim of human parity, and Läubli et al.

(2020, p. 653) argue that the 'finding of human–machine parity was owed to weaknesses in the evaluation design'. Krüger (2022, p. 229) also identifies several biases in 'state-of-the art MT quality evaluation methodologies'. Nonetheless, the claim was widely publicised in the media, giving a perception to the general public that MT is trustworthy.

This sort of broad interpretation of limited-scope evaluation has been criticised by Cartwright (1989, p. 6) who argues that a measuring instrument will only tell us what we wish to learn if it operates on 'the principles that we think it does, and if it is working properly, and if our reading of the output is right'. She warns against broad interpretations of measures and says that a 'measurement that cannot tell us a definite result is no measurement at all' (Cartwright, 1989, p. 6). According to Messick (1989, p. 5), for a measure or score to be considered valid, it should be appropriate, meaningful, and useful, and should also have functional worth 'in terms of the social consequences of their use'. He argues for a unified consideration of validity, whereby appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable from the 'empirical grounding or trustworthiness of the score interpretation' (Messick, 1989, p. 8). According to Sireci (2007, p. 480), validity refers to the *interpretation* of a measure rather than the test itself and can never be unequivocal; thus the 'need to put forth enough evidence to make a convincing argument that the interpretations made on the basis of test scores are useful and appropriate'.

Operationalisation of translation quality is famously difficult, and any measure of quality is likely to be lacking or incomplete (Drugan, 2013). Furthermore, Castilho et al. (2018, p. 30) note that 'a lack of standardisation… has yielded great inconsistency' in quality evaluation. Error-based measures such as multidimensional quality metrics (Lommel, 2018) can help with granular analysis of translation fluency and adequacy or fidelity, although interpretation of categories by evaluators may be inconsistent, particularly without training, leading to poor inter-annotator agreement. In general, MT evaluation metrics tend to be unidimensional. Direct assessment was proposed for competitive shared tasks by Graham et al. (2016) as a replacement for BLEU (Papineni *et al.*, 2002), which had been used previously. Direct assessment, while still a rough measure of accuracy, was found to correlate better with human judgement (as it is based on human judgement!) and to be a more reliable measure for quality in shared task scenarios. As noted by Toral et al. (2018, p. 116), however, crowd evaluators 'tend to be more accepting' of MT than human output. Measures should ideally be aligned to the intended use of MT: instructions should use task-based measurement; MT for post-editing should use measures of post-editing effort; MT for literature should use measures of comprehensibility or narrative engagement (Guerberof-Arenas and Toral, 2022).

There have been longstanding criticisms of the BLEU metric. Callison-Burch et al. (2007) report that it correlates poorly with human judgement, and Kocmi et al. (2021, p. 479) are unequivocal in their recommended best practices for automatic MT evaluation: 'Do not use BLEU', they write, 'it is inferior to other metrics, and it has been overused'. Despite this, BLEU is highly influential in MT development due to it being a fast and cheap yardstick of MT evaluation. The metric is used in neural MT training to identify the convergence point at which training should stop. It is the most widely-used metric in research, often the sole reported measure of MT quality, to the extent that Freitag et al. (2020) and Kocmi et al. (2021) conclude that it has impeded MT development, with developers optimising their systems to maximise BLEU rather than to improve human judgement, particularly in the context of shared tasks. Cartwright (1989, p. 197) identifies a tendency across disciplines to focus exclusively on a single measure, to 'strip away — in our imagination — all that is irrelevant to the concerns of the moment to focus on some single property or set of properties, as if they were separate'. The use of BLEU for MT evaluation is a good example of this tendency. The recommendation by Kocmi et al. (2021) and others is to move to pre-trained neural metrics on the basis that they correlate better with human judgement. Although this brings a risk that systems will just be optimised for one metric rather than another, the use of word embeddings should mean more accurate scoring of synonyms. Nonetheless, pre-trained metrics do not appear to be a panacea, with weaknesses similar to those of neural MT systems identified by Amrhein and Sennrich (2022, p. 1), who write that 'COMET models are not sensitive enough to discrepancies in numbers and named entities'. They additionally 'show that these biases cannot be fully removed by simply training on additional synthetic data' (Amrhein & Sennrich 2022, p. 8).

Schwartz et al. (2020) describe how an exclusive focus on performance when using one MT evaluation metric or another discourages what they term a holistic approach to AI system development. Considerations of efficiency and sustainability are ignored in the quest for ever greater performance, a tendency that pushes AI development in the wrong direction. They advocate research that 'yields novel results while taking into account the computational cost, encouraging a reduction in resources spent' (Schwartz et al. 2020, p. 59). However, the tendency to focus narrowly on performance as represented by a small number of attributes appears to be becoming more rather than less common, with value judgements also applied to the work of translators and many other types of workers with serious repercussions.

Algorithmic Norms in Translation

The previous sections outline problems with quantification of translation and particularly with mathematisation within opaque machine learning systems. This section introduces the concept of algorithmic norms as an effect of mathematisation on translation workers, taking an algorithm as a sequence or routine of mathematical operations on numbers or symbols.

Translation quality evaluation scores can affect translators' reputations within an organisation, but as more translation activity data is gathered during translation, particularly within cloud-based translation platforms, many data points may be used as proxies for evaluation of translator performance. As translators working on these platforms are engaged on a freelance per-project basis, these scores will dictate their chances of receiving further work. The way that these data are combined for evaluation is rarely published, but there is growing evidence of automatic job allocation or limitations to online translation job availability based on previous performance. One company that has been willing to reveal the data points used within their algorithm for translator evaluation is the company Translated. Cattelan (2017) explains that their T-Rank system had (in 2017) been trained on over 980,000 translation jobs. The basis for the T-Rank score is collected translation activity data, including data on the 'type of job, on the domain, on the translation quality, timeliness of the delivery, layout and formatting, communication skills, feedback from the PM' (Cattelan 2017, 11). T-Rank is used for semi-automatic job allocation from over 250,000 translators who have registered on Translated's Matecat platform. Similarly, Massardo (2019) explains Wordbee's Translation Quality Index, which is used for reporting to clients rather than for automated decision-making, and comprises a Capacity Utilisation Ratio based on output assigned and completed within a set time frame, Delivery In Full On Time Rate based on job acceptance and completion time, First Pass Yield Rate based on reliability and adherence to instructions, Order Fulfilment Cycle Time based on a timeliness rating and job completion before deadlines, Rework Level based on a quality rating and the number of segments that require correction, and User Ratings. The quality index is then used as a basis for future job allocation, although Massardo (2019) advises clients particularly to double-check user ratings due to their subjective nature.

In theory, such measures should benefit translators who work and communicate well, and Fairwork (2022) note that they may help to minimise unpaid work in the form of time spent trawling through translation jobs on online platforms. Fairwork (2022, p. 15) report that translators spend on average 3% of their time on the Translated platform on unpaid work, as compared to almost 20% on the Rev platform. However, it is difficult to see how these quality scores could be validated. Messick (1989, 5) is particularly concerned about the use of scoring as a basis for action, and writes that 'what is to be validated is not the test or the observation device as such, but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails'. Cattelan (2017) reports that Translated created a benchmark for job allocation based on the average choices of experienced project and vendor managers and that T-Rank had a 54% match with the benchmark – better than the average project or vendor manager, but less than the best-performing (although we can probably assume that this has since improved). Even if translators benefit from the system, there is a risk that poor scoring due to personal circumstances or illness might lead to reduced offers of work or limited access to jobs. Plenty of authors, such as O'Neil (2016) have highlighted the risks inherent in leaving decision-making solely to an algorithm without any oversight, and even where there is oversight, the high likelihood that human operators will adhere to an automated proposal rather than assuming full liability by ignoring it.

As noted previously, Kocmi *et al.*, (2021) and Schwartz *et al.* (2020) believe that MT and AI development has been shaped by methods of evaluation, whether these are specific metrics such as BLEU or a focus on 'measures of performance such as accuracy, at the expense of measures of efficiency' (Schwartz et al. 2020, p. 57). If human work is also to be algorithmically evaluated, it makes sense that workers will calibrate their work to satisfy the requirements of the algorithm. Adorno and Horkheimer, in 1947 (p. 23), wrote of workers who 'must mold themselves to the technical apparatus, body and soul', as the process of self-preservation 'enforces the self-alienation of individuals'. In response to algorithmic management on gig work platforms, Jarrahi and Sutherland (2019, p. 587) report that the 'ability to understand and make use of algorithms has […] become a core competency of workers attempting to retain autonomy'. If producing work that conforms to algorithmic evaluation is the primary aim for a translation, this represents a sizeable shift from the notion of creating a translation tailored for a particular *skopos* or for an imagined reader and a shift from the *norms* of translation production. The weighting of each element of the algorithmic evaluation in the above examples has not been revealed, but the importance of achieving high scores for speed of responses, communication skills, and adherence to deadlines represent a further shift. These elements were doubtless important when dealing with human project and vendor managers, but were not subject to automatic evaluation.

Toury (2012, p. 63) defines norms as the 'translation of general values or ideas shared by a community […] into performance 'instructions' appropriate for and applicable to concrete situations'. He continues that norms are often not verbalised, but still serve as a barometer for assessment, with rewards for conformity and sanctions or punitive outcomes for non-conformity. These norms are inherently unstable and will differ depending on the time, context, and intended audience. Toury divides norms into a set of preliminary (regarding translation policy and text type) and operational norms (guiding decisions made while translating). Chesterman (1997, p. 56) proposes an extension to Toury's framework of translation norms to include what he calls 'Expectancy Norms'. These norms relate to the extent to which a translation conforms to a reader's expectation for a translated text and allow readers or critics to make an evaluative judgement of a translation. Chesterman (1997, p. 55) explains that these 'stand midway between (judicial) laws and conventions' and may be evaluated by a norm authority, such as a critic or examiner. If the norm authority is an algorithm rather than a human arbiter, this creates a potential mismatch with the needs and expectations of end readers. Toury (2012) discusses competing norms, but his typology of mainstream, old-fashioned and avant-garde norms is dependent on the place of a (literary) text within the culture. For Chesterman, translation competence involves learning about translation strategies and norms. He adds that norms may act as constraints as well as guidelines, and that there ‘may be situations where there is a clash between the norms sanctioned by these norm-authorities and the norms accepted and current in the society at large’ (Chesterman, 1997, p. 66). Algorithmic management introduces norms regarding translation and related communications that differ from those previously identified in that they are not based on literary translation, not calibrated to the audience, they are static and unchanging unless the algorithm is adjusted, and they are automatically and rigorously applied. They are not the same as expectancy norms relating to the expectations of the client or reader based on previous norm-setting translation; they are not quite professional norms although optimising communication is likely to be a key part of both professional and algorithmic norms. They also differ from translation laws in that they cannot be generalised as rules (as described by Olohan (2020)) as they are not codified, although they are institutionally applied. For translators who do not adhere to these algorithmic norms, the punishment will be fewer jobs or restricted access to jobs. They require that translators gain what Jarrahi and Sutherland (2019) refer to as algorithmic competence, reverse engineering the opaque routine of the algorithm.

The conceptualisation of algorithmic norms is important, as tailoring a translation process and product to a set of automatically evaluated requirements may be considered acceptable by the end readers or users and may fit with the professional norms regarding accountability, communication, and relations between source and target text as defined by Chesterman, but it may also clash with those norms – or at the very least encourage a focus on the specific metrics valued by the algorithm at the expense of others. If there is a mismatch between algorithmic norms and expectancy norms, the former are likely to assume more importance to the translators than either expectancy or the subordinate professional norms that aim to maximise the communicative efficiency of a translation. After all, conforming to algorithmic norms will dictate whether a freelance translator will be engaged again or not, and thus fulfilling the algorithmic requirements becomes a primary need.

Sakamoto (2018) notes that algorithmic management cannot hope to include project managers’ ‘tacit knowledge’ (knowledge that is unwritten and not taught) of translators, both positive and negative, based on experience. Herbert et al. (2023) also find that project managers are unhappy at the thought of losing control of job assignment for reasons of job satisfaction and distrust of the algorithm. There has been little research on the effects of algorithmic management on translators to date. The survey by Herbert et al. was carried out in advance of a move to full automation at the unnamed partner company. As with many other automated job allocation systems, this will involve an automatic ranking of translators for each job based on previous work on the chosen language pair, domain, and other factors included in the algorithm (and the translators’ adherence to them). The translators are thus assigned a comparative value and will be contacted in order without human intervention.

‘Six of one, half a dozen of the other’: On mathematisation and systems

Numbers form the basis of quantification, transactions, and communication in translation and are helpful and indisputably necessary. An analysis of quantification in translation that dwells only on negative impacts risks being dismissed as mere ‘opposition to bad things’ (Pym, 2017, p. 367). The intention of this article is to introduce contributions from other fields that might help with ethical and valid decision-making when using numbers and measures in context. Narayanan (2022, p. 17) writes that numbers ‘have been the language of policy making for more than a century, but especially so today, with the tech industry being so successful at convincing the public about the power of big data and AI’. As Vincent (2022) writes, measurement reinforces what we consider important in life. The choice of what to measure is therefore a powerful one, as is the interpretation of that measurement, and its use as a basis for action. While the translation industry could not function without quantification and, increasingly, mathematisation, there are limits to what we can validly infer from scores and measures. Messick (1989, p. 5) proposes four interrelated questions to pose when considering validity:

> 1. What balance of evidence supports the interpretation of meaning of the scores?
>
> 2. What evidence undergirds not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting?
>
> 3. What rationales make credible the value implications of the score interpretation and any associated implications for action?
>
> 4. What evidence and arguments signify the functional worth of the testing in terms of its intended and unintended consequences?

The question raised by Habermas (1984/2015, p. 115) then, is 'who decides what is valid?', as the receiver of information has to understand the conditions and context of a validity claim. According to Habermas' (1984/2015, 66) theory of communicative action, there are limitations to what can be empirically understood as true, and some knowledge may be 'insightfully recapitulated from the perspective of participants'. Thus, validity and acceptance of a judgement should, he believes, be achieved through communication and understanding. While it may not be possible or desired to achieve communicative understanding on every topic – sometimes knowledge might be assumed or the numbers themselves may be sufficient – Habermas (1987, p. 298) diagnoses a steady colonisation of the social, cultural, and personal by the strictly rational, culminating in a 'technicising' process whereby kindness and fairness are replaced by means-ends rationality subject to 'the imperatives of autonomous subsystems'. While acknowledging that they reduce unpaid work as previously discussed, this does seem to describe the algorithmic approach to automation or semi-automation of project management steps – so-called 'lights-out translation project management'.

Lights-out project management follows the logic of our techno-social aims, as identified by Frischmann and Benesch (2023, p. 387), to 'maximise efficiency, minimise transaction costs, eliminate friction, seamlessly interconnect, and increase the speed, scale and scope of engineered interactions'. They call such techno-social management 'superficially defensible yet deeply flawed upon examination' as there is a need to engineer in some friction for governance, oversight, and to set boundaries. Friction allows for reflection and self-determination. The point here is not to accept the default without reflection, as 'we want neither to eliminate friction nor to have too much of it' (Frischmann and Benesch, 2023, p. 389). Unfortunately, the hype about AI in the media following the public launch of generative tools is unlikely to encourage what Heidegger (1966) called meditative thinking. This may be why the translators, company representatives, academics and students who responded to the 2023 ELIS survey perceive AI as 'a negative trend', reversing their positive perception of the previous year (ELIS Research, 2023, p. 4).

The opacity of subsymbolic neural networks for both translation and job allocation is a further reason to take care with the use and interpretation of their outputs. Duede (2022, p. 491) believes that our confidence in neural outputs is often based on the 'consistent success of that process in producing accurate results'. However, because each input is previously unseen and untested, we cannot know for certain whether this input is a good fit for our system based on its training data and cannot see how the 'individual parts interact and contribute to the network's outputs' (Duede, 2022, ibid.). This does not mean that the outputs are not excellent or useful, just that the reliability of the system is uncertain due to the opacity of the process, thus the regular recommendation from Way (2018) and others that the degree of automation of translation should be appropriate for the shelf-life, risk, and value of a text. It follows that decisions on job allocation should be monitored and audited.

As described previously by Dunne (2012) and others, the context in the translation industry is very often one of information asymmetry, in which translators feel alienated from their work as what Lukács, Livingstone and Lukács (2013, p. 89) call ‘a mechanical part incorporated into a mechanical system’, under constant surveillance. The industry is multifarious, and this is not the experience of all translators. However, for those working within highly restrictive platforms, subject to opaque algorithmic evaluation, there is a frustration and a diminution of agency when confronted with a faceless mathematised process as a basis for inscrutable decisions. Our efforts to maximise validity should particularly protect those most vulnerable. Following the view that ‘society is not determined by technology, nor is technology determined by society’ (Bijker, 1999, p. 274), good intentions are not enough to make ethical decisions on the use of quantification and mathematisation in translation processes, nor will a deontological set of rules suffice, although there is a place for legislation to set legal limitations on data processing and inferences (as is increasingly necessary as algorithmic decision-making presents risks in many areas of work). As an example of ongoing efforts to set legal limits, the European Union’s draft AI Act (European Commission, 2021, p. 26) designates systems for ‘task allocation, monitoring or evaluation of persons in work-related contractual relationships’ as high risk, requiring retention of documentation, conformity assessments, transparency and, importantly, human oversight.

There is a degree of polarisation of positions regarding unfettered scientific progress and the proposals for legal and ethical limits to AI. The achievements of AI systems are incredible and impressive, but while not all dangers highlighted by researchers in AI ethics are critical, they should not be ignored and should be considered in the context of risk. Our challenge is to find a way for systems’ properties and our values to ‘work together to bring forth something much better than could ever be produced by our will alone’ (Meadows, 2008, p. 170). A focus on sustainable work systems (Docherty, Kira and Shani, 2008) will require a continuous process of evaluation and reevaluation of workflow steps and decisions in order to prioritise long-term benefits to work system stakeholders over short-term gains. We might then ask: who are these stakeholders? The traditional stakeholder approach (following Phillips, 2003) from business ethics allows for flexibility in analysing possible consequences of quantification and may be more readily accepted by those with power in the industry, as by design it prioritises central actors, whereas an ethics of care approach (Held, 2005) would instead prioritise the most vulnerable. Stakeholder identification notwithstanding, a key and actionable recommendation for sustainable translation work is for transparency in the use of algorithmic decision-making, reversing the trend towards opacity that tends to disempower workers, particularly within contemporary translation platforms. This does not mean eschewing quantification and mathematisation, but rather describing their use and basis in data clearly. Such transparency will not benefit those who wish to ignore unfair practices within translation production, but will assist translators and translation buyers in their choices while helping to assert the value of translation.

Acknowledgement

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/2106_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Amrhein, C. and Sennrich, R. (2022) ‘Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET’. arXiv. Available at: https://doi.org/10.48550/ARXIV.2202.05148.

Bapna, A. *et al.* (2022) ‘Building machine translation systems for the next thousand languages’. arXiv. Available at: https://doi.org/10.48550/arXiv.2205.03983.

Berghofer, P., Goyal, P. and Wiltsche, H.A. (2021) ‘Husserl, the mathematization of nature, and the informational reconstruction of quantum theory’, *Continental Philosophy Review*, 54(4), pp. 413–436. Available at: https://doi.org/10.1007/s11007-020-09523-8.

Bijker, W.E. (1999) *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*. 3. Aufl. Cambridge: MIT Pr (Inside technology).

Callison-Burch, C. *et al.* (2007) ‘(Meta-) Evaluation of Machine Translation’, in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, pp. 136–158. Available at: https://aclanthology.org/W07-0718.

Cartwright, Nancy. (1989) *Nature’s capacities and their measurement Nancy Cartwright.* Oxford: Clarendon Press (Clarendon Paperbacks Ser.).

Castilho, S. *et al.* (2018) ‘Approaches to Human and Machine Translation Quality Assessment’, in J. Moorkens et al. (eds) *Translation Quality Assessment*. Cham: Springer International Publishing (Machine Translation: Technologies and Applications), pp. 9–38. Available at: https://doi.org/10.1007/978-3-319-91241-7_2.

Castilho, S. *et al.* (2023) ‘Do online machine translation systems care for context? What about a GPT model?’, in. *24th Annual Conference of the European Association for Machine Translation (EAMT 2023)*, Tampere, Finland. Available at: https://doras.dcu.ie/28297/.

Cattelan, A. (2017) ‘T-Rank. A decision support system for project assignment’. *META-FORUM 2017*, Brussels, Belgium, 13 November. Available at: http://www.meta-net.eu/events/meta-forum-2017/pdf/02-Session-06-06-Alessandro-Cattelan.pdf.

Chaitin, G.J. (2013) *Proving darwin: making biology mathematical*. 1st Vintage books ed. New York: Vintage Books.

Chesterman, A. (1997) *Memes of translation: the spread of ideas in translation theory*. Amsterdam: John Benjamins.

Chomsky, N. (1956) ‘Three Models for the Description of Language’, *IRE Transactions on Information Theory*, 2, pp. 113–124.

Docherty, P., Kira, M. and Shani, A.B. (2008) ‘What the world needs now is sustainable work systems’, in P. Docherty, M. Kira, and A.B. Shani (eds) *Creating sustainable work systems: developing social sustainability*. 2. ed. London New York, NY: Routledge.

Drugan, J. (2013) *Quality in professional translation: assessment and improvement*. Bloomsbury. Available at: https://www.bloomsbury.com/uk/quality-in-professional-translation-9781441149541/ (Accessed: 19 September 2022).

Duede, E. (2022) ‘Instruments, agents, and artificial intelligence: novel epistemic categories of reliability’, *Synthese*, 200(6), p. 491. Available at: https://doi.org/10.1007/s11229-022-03975-6.

Duhem, P.M.M. (1954) *The aim and structure of physical theory*. Princeton: Princeton University Press.

Dunne, K.J. (2012) ‘The industrialization of translation: Causes, consequences and challenges’, *Translation Spaces*, 1, pp. 143–168. Available at: https://doi.org/10.1075/ts.1.07dun.

ELIS Research (2023) *European language industry survey 2023: Trends, expectations and concerns of the European language industry*. 10. ELIA, EMT, EUATC, FIT Europe, GALA, LIND, Women In Localization, pp. 1–56. Available at: https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf.

Ellul, J. (1964) *The technological society: a penetrating analysis of our technical civilization and of the effect of an increasingly standardized culture on the future of man*. [Nachdruck der Ausgabe] New York, Knopf. Translated by J. Wilkinson. New York, NY: Vintage books (A Vintage book).

European Commission (2021) *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Fairwork (2022) *Working Conditions in the Global Platform Economy: Fairwork Translation and Transcription Platform Ratings 2022*. Oxford, UK: Fairwork.

Fırat, G., Gough, J. and Moorkens, J. (2024) ‘Working Conditions of Translation Workers in the Digital Platform Economy’, *Perspectives* [Preprint], (forthcoming).

Folaron, D.A. (2012) ‘Digitalizing translation’, *Translation Spaces*, 1, pp. 5–31. Available at: https://doi.org/10.1075/ts.1.02fol.

Freitag, M., Grangier, D. and Caswell, I. (2020) ‘BLEU might be Guilty but References are not Innocent’, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 61–71. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.5.

Frischmann, B.M. and Benesch, S. (2023) ‘Friction-In-Design Regulation as 21St Century Time, Place and Manner Restriction’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.4178647.

Golumbia, David. (2009) *The cultural logic of computation David Golumbia.* Cambridge, Mass: Harvard University Press. Available at: https://doi.org/10.4159/9780674053885.

Graham, Y. *et al.* (2016) ‘Is all that Glitters in Machine Translation Quality Estimation really Gold?’, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3124–3134. Available at: https://aclanthology.org/C16-1294.

Guerberof-Arenas, A. and Toral, A. (2022) ‘Creativity in translation: Machine translation as a constraint for literary texts’, *Translation Spaces*, 11(2), pp. 184–212. Available at: https://doi.org/10.1075/ts.21025.gue.

Guerreiro, N.M., Voita, E. and Martins, A.F.T. (2022) 'Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation'. arXiv. Available at: http://arxiv.org/abs/2208.05309 (Accessed: 21 November 2022).

Habermas, J. (1984) *The theory of communicative action: 1. Reason and the rationalization of society*. Place of publication not identified: Polity Press.

Habermas, J. (1987) *The theory of communicative action: 2. Lifeworld and systems, a critique of functionalist reason*. Place of publication not identified: Polity Press.

Hassan, H. *et al.* (2018) 'Achieving Human Parity on Automatic Chinese to English News Translation'. arXiv. Available at: http://arxiv.org/abs/1803.05567 (Accessed: 6 December 2022).

Heidegger, M. (1966) *Discourse on Thinking*. Translated by J.M. Anderson and E.H. Freund. Harper & Row (Harper colophon books). Available at: https://books.google.ie/books?id=y8k7AQAAIAAJ.

Held, V. (2005) *The Ethics of Care: Personal, Political, and Global*. 1st edn. Oxford University PressNew York. Available at: https://doi.org/10.1093/0195180992.001.0001.

Henderson, L. (2022) 'The Problem of Induction', in E.N. Zalta and U. Nodelman (eds) *The Stanford Encyclopedia of Philosophy*. Winter 2022. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2022/entries/induction-problem/.

Hendy, A. *et al.* (2023) 'How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation'. arXiv. Available at: https://doi.org/10.48550/ARXIV.2302.09210.

Herbert, S. *et al.* (2023) 'From responsibilities to responsibility: a study of the effects of translation workflow automation', *Journal of Specialised Translation* [Preprint], (40).

Horkheimer, M. and Adorno, T.W. (1947) *Dialectic of enlightenment: philosophical fragments*. Edited by G. Schmid Noerr. Translated by E. Jephcott. Stanford, Calif: Stanford University Press (Cultural memory in the present).

International Organization for Standardization (no date) 'ISO 18587:2017'. Available at: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/29/62970.html (Accessed: 2 May 2022).

Jarrahi, M.H. and Sutherland, W. (2019) 'Algorithmic Management and Algorithmic Competencies: Understanding and Appropriating Algorithms in Gig Work', in N.G. Taylor et al. (eds) *Information in Contemporary Society*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 578–589. Available at: https://doi.org/10.1007/978-3-030-15742-5_55.

Kocmi, T. *et al.* (2021) 'To ship or not to ship: An extensive evaluation of automatic metrics for machine translation', in *Proceedings of the Sixth Conference on Machine Translation*. *EMNLP-WMT 2021*, Online: Association for Computational Linguistics, pp. 478–494. Available at: https://aclanthology.org/2021.wmt-1.57 (Accessed: 2 May 2022).

Krüger, R. (2022) 'Some Translation Studies informed suggestions for further balancing methodologies for machine translation quality evaluation', *Translation Spaces*, 11(2), pp. 213–233. Available at: https://doi.org/10.1075/ts.21026.kru.

Läubli, S. *et al.* (2020) ‘A Set of Recommendations for Assessing Human–Machine Parity in Language Translation’, *Journal of Artificial Intelligence Research*, 67. Available at: https://doi.org/10.1613/jair.1.11371.

Levitina, N. (2011) ‘Requirements collection: The foundation of scope definition and scope management in localization projects’, in K.J. Dunne and E.S. Dunne (eds) *American Translators Association Scholarly Monograph Series*. Amsterdam: John Benjamins Publishing Company, pp. 95–118. Available at: https://doi.org/10.1075/ata.xvi.07lev.

Lommel, A. (2018) ‘Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies’, in J. Moorkens et al. (eds) *Translation Quality Assessment*. Cham: Springer International Publishing (Machine Translation: Technologies and Applications), pp. 109–127. Available at: https://doi.org/10.1007/978-3-319-91241-7_6.

Lukács, G., Livingstone, R. and Lukács, G. (2013) *History and class consciousness: studies in Marxist dialects*. Nachdr. Cambridge, Mass: MIT Press.

Marais, K. and Meylaerts, R. (2018) ‘Introduction’, in K. Marais and R. Meylaerts (eds) *Complexity Thinking in Translation Studies: Methodological Considerations*. 1st edn. Routledge, pp. 1–18. Available at: https://doi.org/10.4324/9780203702017.

Marcus, G. (2020) ‘The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence’. arXiv. Available at: http://arxiv.org/abs/2002.06177 (Accessed: 22 November 2022).

Massardo, I. (2019) ‘Business Analytics for Translation and Localization: The Quality Index’, *Wordbee*, 10 January. Available at: https://wordbee.com/blog/localization-industry/business-analytics-for-translation-and-localization-the-quality-index/ (Accessed: 1 December 2022).

Meadows, D. (2008) *Thinking in Systems*. Edited by D. Wright. Chelsea Green Publishing.

Mellinger, C.D. and Hanson, T.A. (2017) *Quantitative research methods in translation and interpreting studies*. London ; New York: Routledge.

Messick, S. (1989) ‘Meaning and Values in Test Validation: The Science and Ethics of Assessment’, *Educational Researcher*, 18(2), pp. 5–11. Available at: https://doi.org/10.3102/0013189X018002005.

Mitchell, M. (2020) *Artificial intelligence: a guide for thinking humans*. Published in paperback. London: Pelican, an imprint of Penguin Books (A Pelican book).

Narayanan, A. (2022) ‘The limits of the quantitative approach to discrimination’. *James Baldwin lecture*, Princeton University, 11 October. Available at: https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/.

NLLB team *et al.* (2022) ‘No language left behind: Scaling human-centered machine translation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2207.04672.

Olohan, M. (2020) *Translation and practice theory*. London ; New York: Routledge (Translation theories explored).

O’Neil, C. (2016) *Weapons of math destruction: how big data increases inequality and threatens democracy*. First edition. New York: Crown.

Papineni, K. *et al.* (2002) ‘Bleu: A method for automatic evaluation of machine translation’, in P. Isabelle, E. Charniak, and D. Lin (eds) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. *ACL 2002*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. Available at: https://doi.org/10.3115/1073083.1073135.

Phillips, R. (2003) *Stakeholder theory and organizational ethics*. 1st ed. San Francisco: Berrett-Koehler.

Piper, A. (2018) *Enumerations: data and literary study*. Chicago ; London: The University of Chicago Press.

Popel, M. *et al.* (2020) ‘Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals’, *Nature Communications*, 11(1), p. 4381. Available at: https://doi.org/10.1038/s41467-020-18073-9.

Pym, A. (2017) ‘Translation and economics: inclusive communication or language diversity?’, *Perspectives*, 25(3), pp. 362–377. Available at: https://doi.org/10.1080/0907676X.2017.1287208.

Rudin, C. (2019) ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nature Machine Intelligence*, 1(5), pp. 206–215. Available at: https://doi.org/10.1038/s42256-019-0048-x.

Sakamoto, A. (2018) ‘Disruption in translator-client matching : paid crowdsourcing platforms vs human project managers’, *Revista Tradumàtica*, (16), pp. 85–94. Available at: https://doi.org/10.5565/rev/tradumatica.218.

Schwartz, R. *et al.* (2020) ‘Green AI’, *Communications of the ACM*, 63(12), pp. 54–63. Available at: https://doi.org/10.1145/3381831.

Shannon, C.E. (1948) ‘A Mathematical Theory of Communication’, *Bell System Technical Journal*, (27), pp. 379–423.

Sireci, S.G. (2007) ‘On Validity Theory and Test Validation’, *Educational Researcher*, 36(8), pp. 477–481. Available at: https://doi.org/10.3102/0013189X07311609.

Skovsmose, O. (2020) ‘Mathematization as Social Process’, in S. Lerman (ed.) *Encyclopedia of Mathematics Education*. Cham: Springer International Publishing, pp. 605–608. Available at: https://doi.org/10.1007/978-3-030-15789-0_112.

Toral, A. *et al.* (2018) ‘Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation’, in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 113–123. Available at: https://doi.org/10.18653/v1/W18-6312.

Toury, G. (2012) *Descriptive translation studies - and beyond*. Rev. ed., 2. expanded ed. Amsterdam: Benjamins (Benjamins translation library, 100).

Van Harmelen, F. (2022) ‘Preface’, in P. Hitzler and M. Kamruzzaman Sarker (eds) *Neuro-Symbolic Artificial Intelligence: The State of the Art*. Amsterdam, NL: IOS Press (Frontiers in Artificial Intelligence and Applications), pp. i–xii.

Vanmassenhove, E., Emmery, C. and Shterionov, D. (2021) 'NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives'. arXiv. Available at: http://arxiv.org/abs/2109.06105 (Accessed: 22 November 2022).

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (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), 1515-1532. https://doi.org/10.1080/1369118X.2020.1776370

Vincent, J. (2022) *Beyond measure: the hidden history of measurement from cubits to quantum constants*. First American edition. London, UK: Faber & Faber.

Way, A. (2010) 'Panning for EBMT gold, or "remembering not to forget"', *Machine Translation*, 24(3), pp. 177–208. Available at: https://doi.org/10.1007/s10590-010-9085-2.

Way, A. (2018) 'Quality expectations of machine translation', in J. Moorkens et al. (eds) *Translation Quality Assessment: From Principles to Practice*. Available at: http://arxiv.org/abs/1803.08409 (Accessed: 2 May 2022).

Weaver, W. (1949) 'Translation', in A.D. Booth and W.N. Locke (eds) *Machine translation of languages: fourteen essays*, pp. 15–23.

Zanettin, F. (2013) 'Corpus Methods for Descriptive Translation Studies', *Procedia - Social and Behavioral Sciences*, 95, pp. 20–32. Available at: https://doi.org/10.1016/j.sbspro.2013.10.618.

Zydroń, A. (2014) 'GMX-V: Slaying the word count dragon', *Multilingual*, 29 June, pp. 33–36.

Zydroń, A. (2017) 'GMX-V Localization Industry Word Count Standard'. *Translatig & the Computer 39*, London, UK, November.