DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Machine Translation Quality Assessment Lesson 1 - Using Metrics Critically

M.Ed. João Lucas Cavalheiro Camargo

Learning Outcomes

LO1 – Develop the awareness of the role of Human Evaluation (HE) in the development of Machine Translation (MT) systems.

LO2 – Develop the awareness of the role of Automatic Evaluation Metrics (AEM) in the development of MT systems.

LO3 – Develop the understanding of key elements in the developments of MT systems, including purpose of system, evaluation metrics, type of evaluators and ethical considerations, to ensure systems are less risky and less biased.

Structure

1 - What do you know about quality assessment?

2 - Quality – Humans and Machines

3 - Purpose of evaluating a MT system

4 - The AI hype

5 - Evaluation – Where do I start?

6 - Metrics

7 - Inter-Annotator Agreement

What do you know about quality assessment?

What do you know about quality assessment?

Assessment

Academia and Industry put effort in assessing the quality of machine translation (MT) systems

(Way, 2018)

Type of Assessment

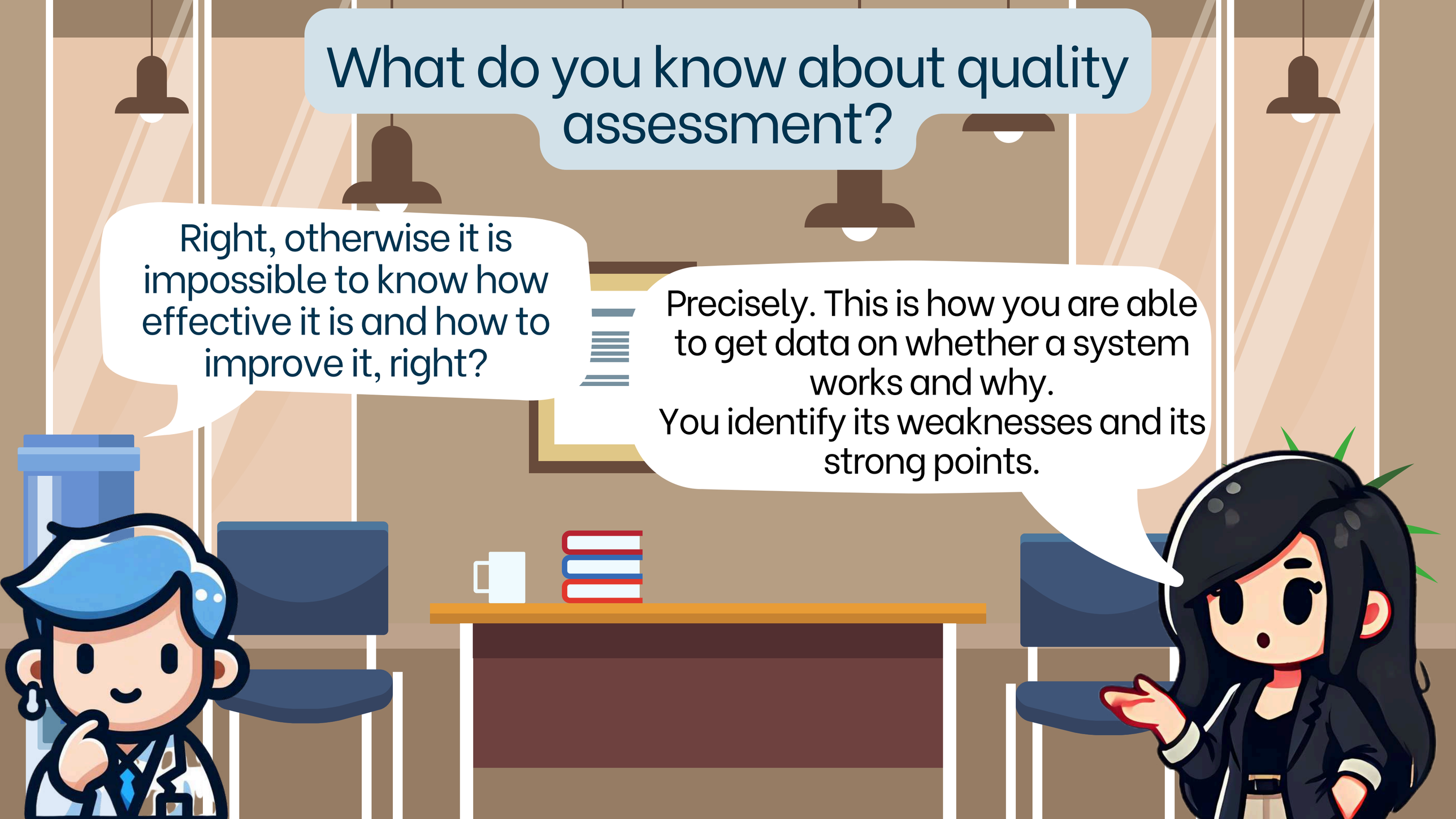
Typically involves automatic evaluation metrics (AEMs) or human evaluation (HE)

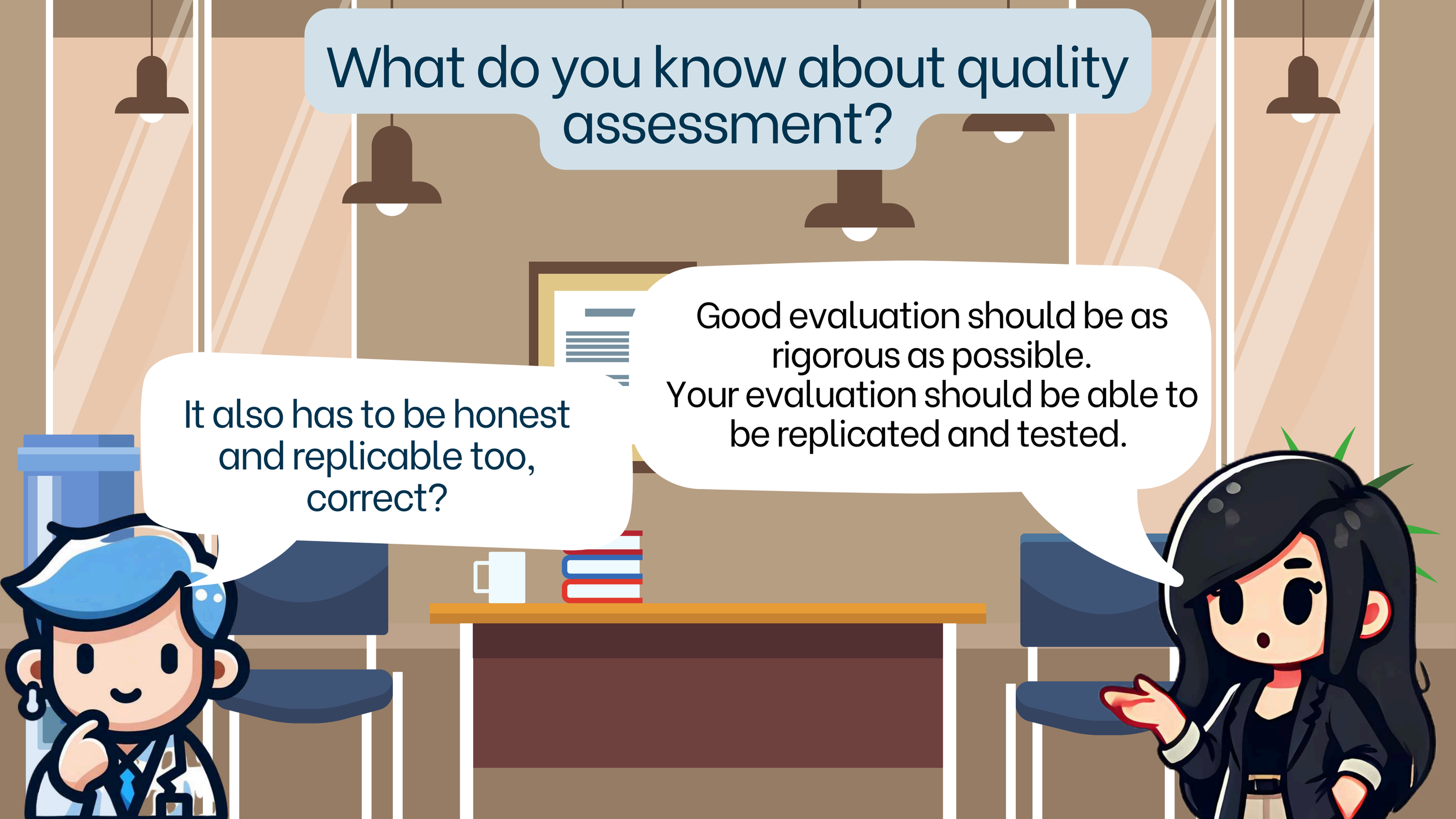
(Castilho *et al.*, 2018)

What do you know about quality assessment?

Why exactly do we need evaluation?

Let me explain! How do you know if a system works?
You evaluate it. So you can compare and report its performance.


What do you know about quality assessment?
Right, otherwise it is impossible to know how effective it is and how to improve it, right?
Precisely. This is how you are able to get data on whether a system works and why.
You identify its weaknesses and its strong points.


What do you know about quality assessment?
It also has to be honest and replicable too, correct?
Good evaluation should be as rigorous as possible.
Your evaluation should be able to be replicated and tested.

Quality – Humans and Machines

AEM
Uses

- Cost-effective
- Quick check for quality improvement

AEM
Drawbacks

- Some may not correlate well with human judgements
- May not be task appropriate
- Are not able to identify detailed insights about translation errors

HE
Uses

- Gold standard
- Offers nuanced insights on translation errors
- Identifies strengths and weaknesses of systems
- Aids in identifying bias in systems

HE
Drawbacks

- Higher costs
- Time-consuming

(Castilho, 2023)

Quality – Humans and Machines

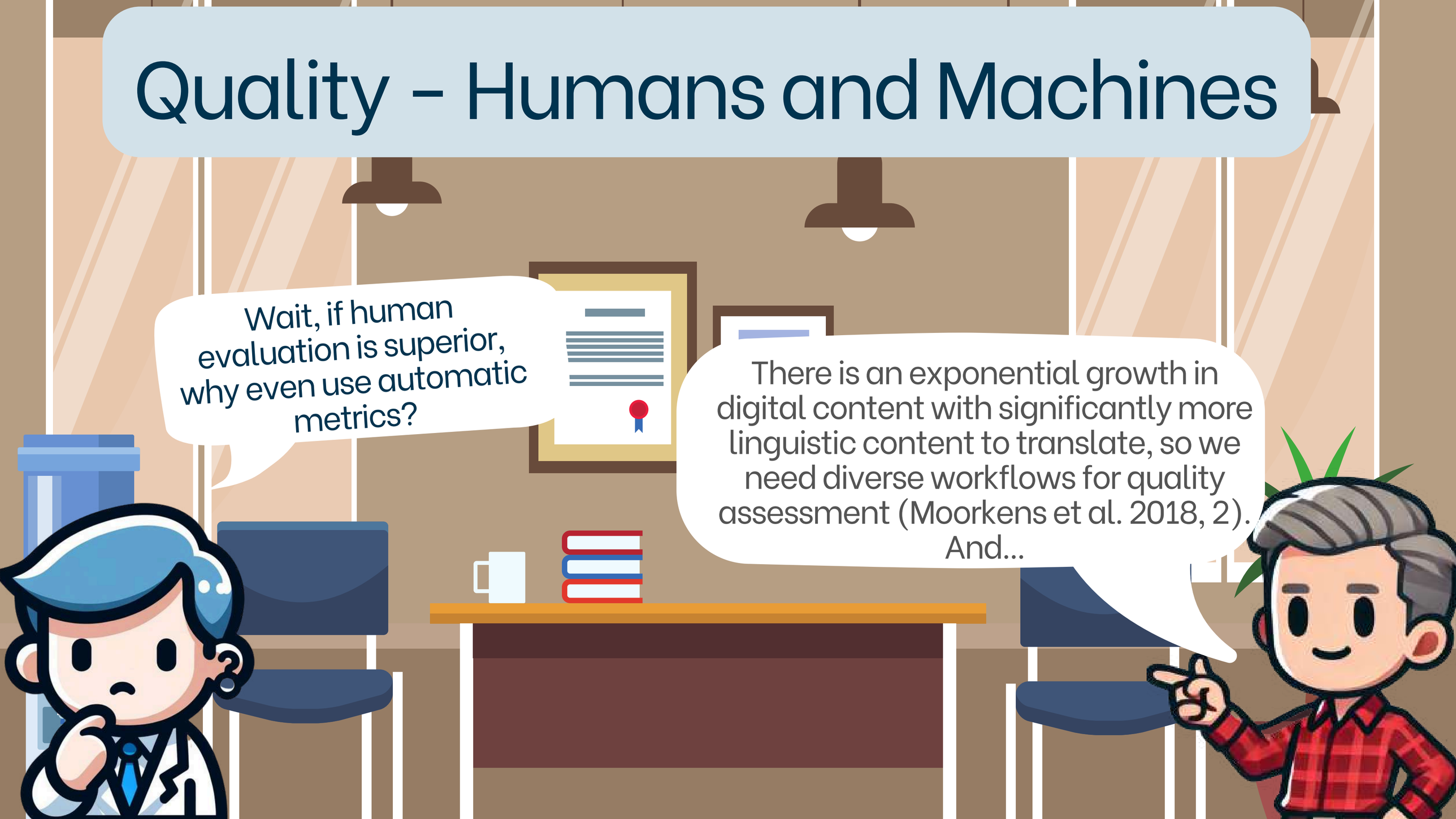
Human Evaluation

- Measures quality
- Provides feedback with detail
- Understands complex linguistic phenomena

Automatic Evaluation

- Can’t measure quality – measures similarity.
- Either needs a reference or needs to be trained on data.

Quality – Humans and Machines



Quality – Humans and Machines

Thinking of it as a workflow:
You use automatic evaluation as a baseline, then you confirm results or investigate deeper with human evaluation.

Human evaluation and automatic evaluation work together, never alone.

Purpose of evaluating a MT system

- Different users, different needs

Type of user	Type of task	Purpose
Translator	Post-editing	Productivity? Satisfaction?
End-user	Comprehension	Gisting? Dissemination?
Both	Diagnosis	Improvement? Detecting flaws?

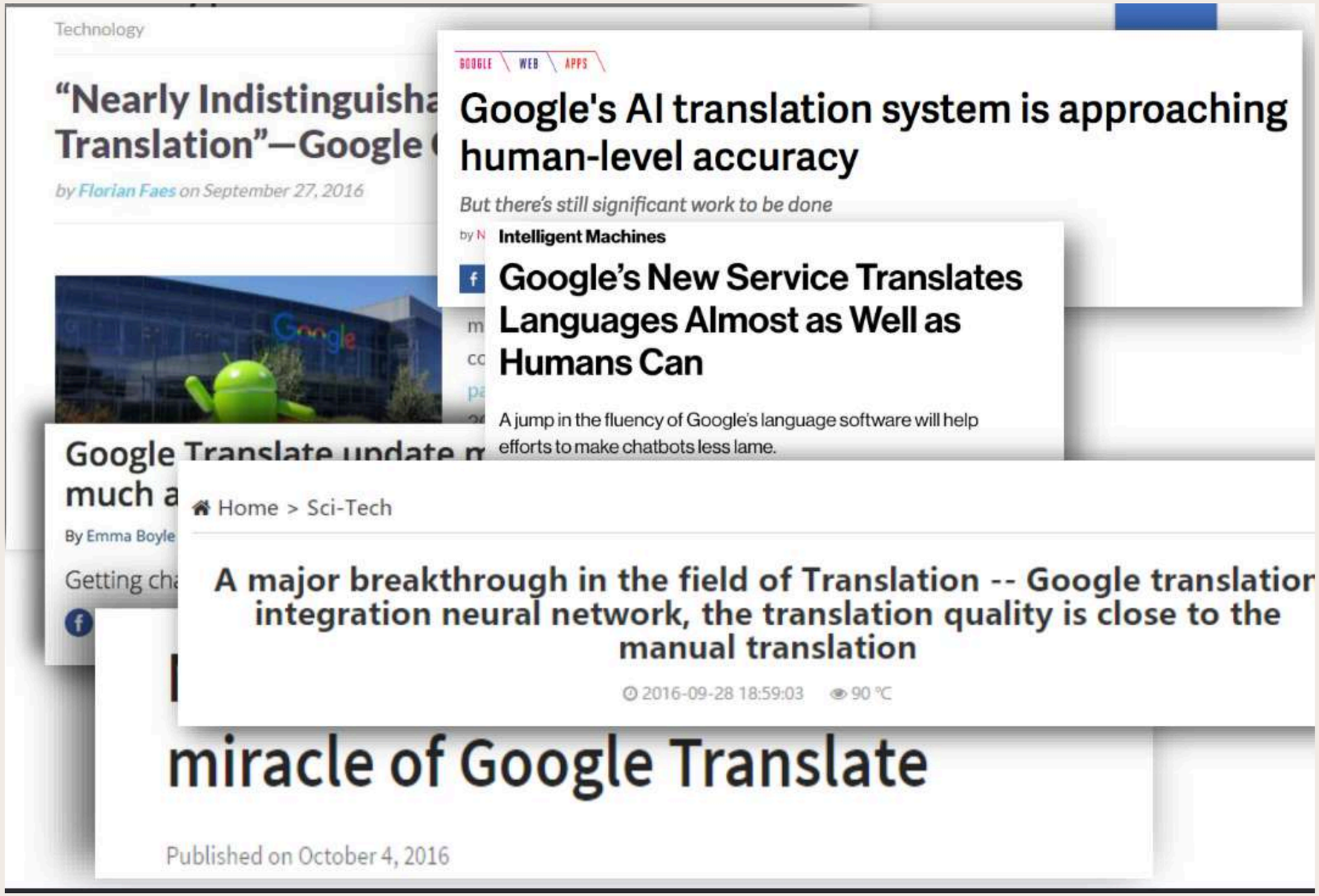
Purpose of evaluating a MT system

Goals

- **Meaningful:** Results should give intuitive interpretation of translation quality
- **Replicable:** Repeating the methodology should lead to the same results
- **Correct:** Metrics must rank better systems higher and must be used appropriately.
- **Low cost:** Reduce time and money spent to carry out evaluation.

Further training can make workflows faster and make them cost less. (Doherty et al. 2018)

The AI Hype


Technology
"Nearly Indistinguisha
Translation"—Google
by Florian Faes on September 27, 2016
GOOGLE WEB APPS
Google's AI translation system is approaching human-level accuracy
But there's still significant work to be done
Intelligent Machines
Google's New Service Translates Languages Almost as Well as Humans Can
A jump in the fluency of Google's language software will help efforts to make chatbots less lame.
Google Translate update
much a
By Emma Boyle
Getting cha
Home > Sci-Tech
A major breakthrough in the field of Translation -- Google translation integration neural network, the translation quality is close to the manual translation
2016-09-28 18:59:03
90 ℃
miracle of Google Translate
Published on October 4, 2016

The AI Hype

Microsoft reaches a historic milestone, using AI to match human performance in translating n

March 14, 2018 | Allison Linn

Huge breakthrough: Microsoft's new AI translates Chinese to English as good as humans

Micr
tran:

Microsoft says its AI can translate Chinese as well as a human

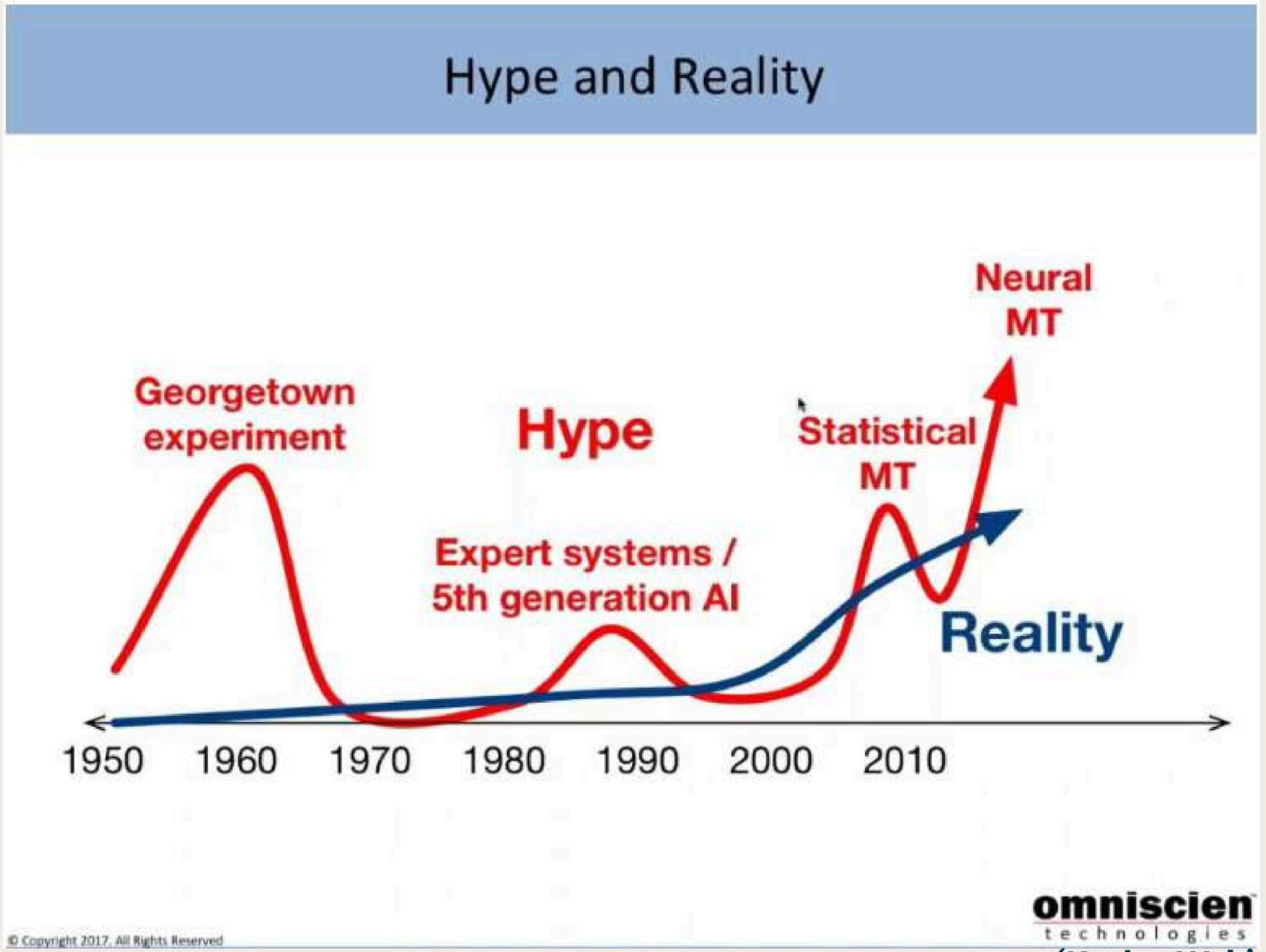
It would represent a major milestone in language-savvy AI.

Technology

'Human Parity Achieved' in Machine Translation — Unpacking Microsoft's Claim

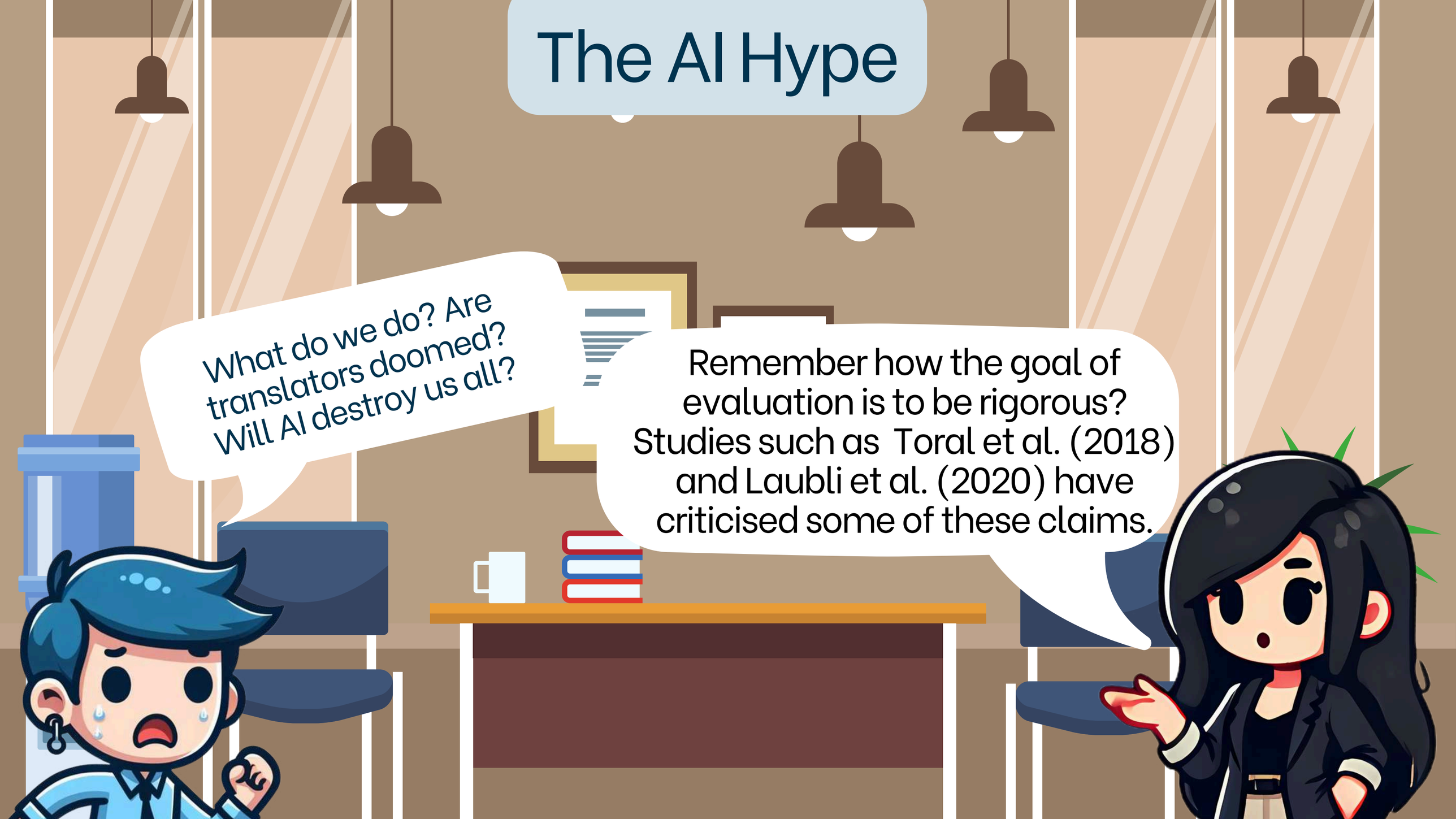
by Gino Diño on March 15, 2018

The AI Hype

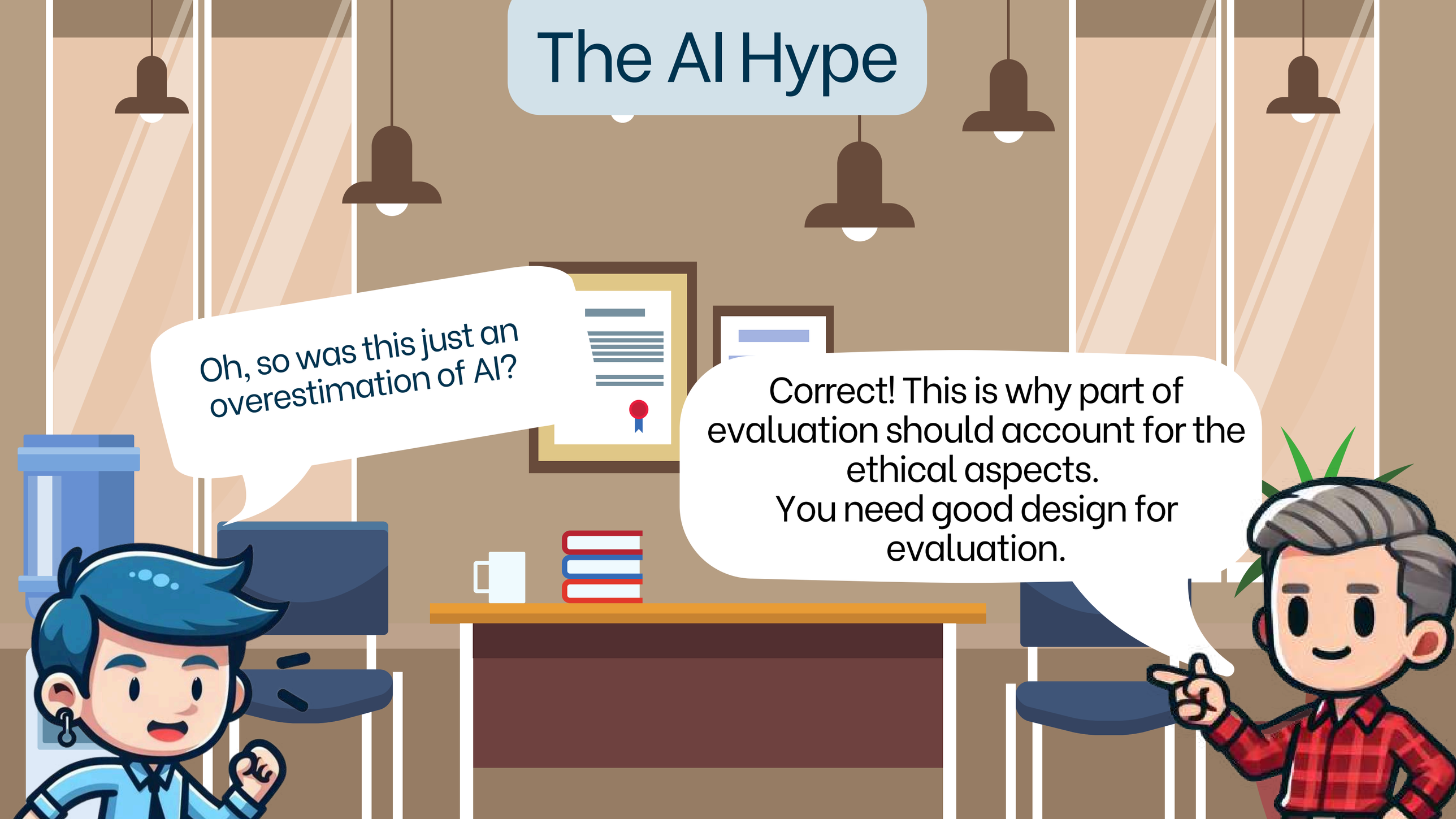


(Koehn, Webinar by Omniscien, 2017)

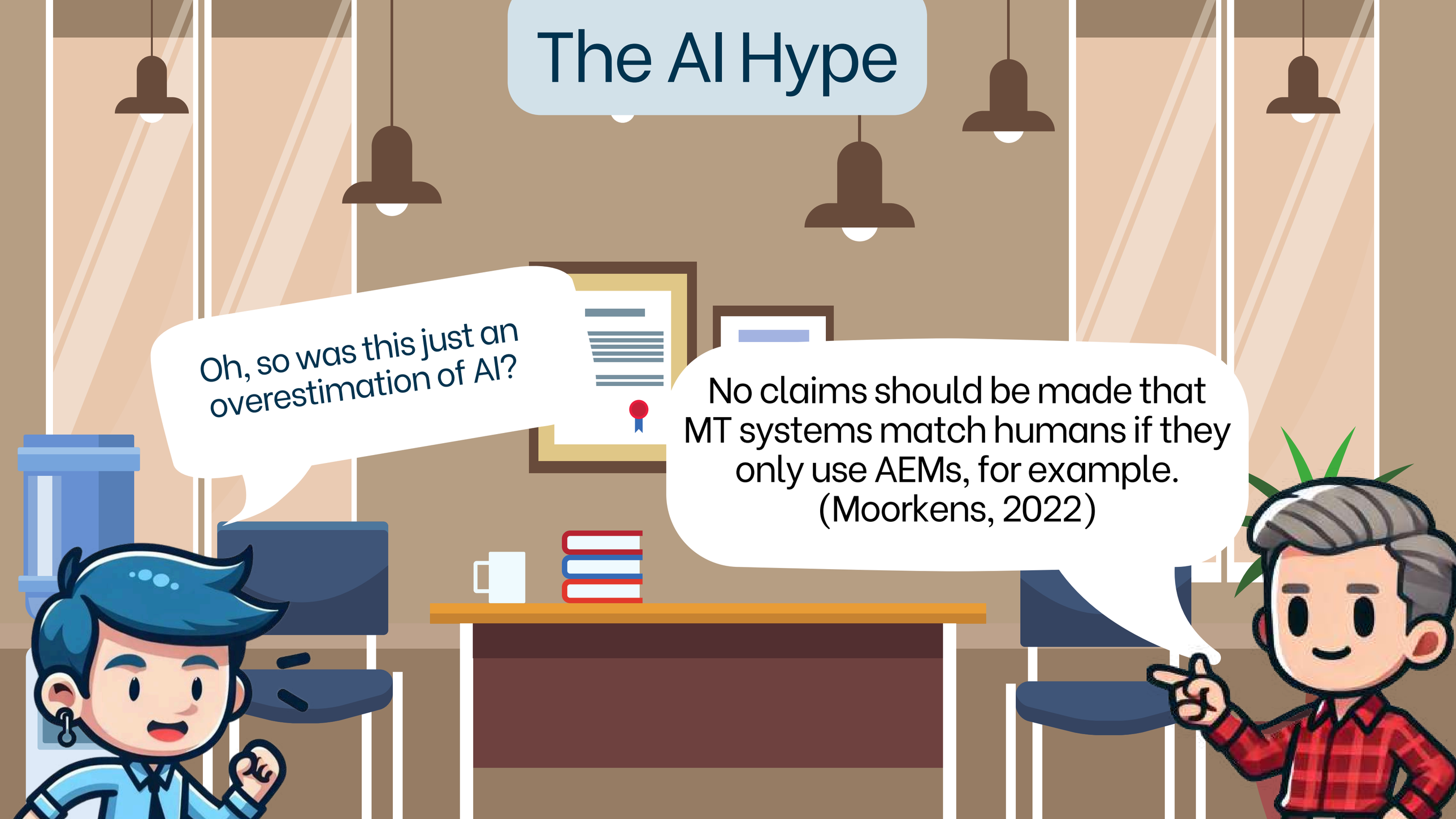
The AI Hype


What do we do? Are translators doomed? Will AI destroy us all?
Remember how the goal of evaluation is to be rigorous? Studies such as Toral et al. (2018) and Laubli et al. (2020) have criticised some of these claims.

The AI Hype


Oh, so was this just an overestimation of AI?
Correct! This is why part of evaluation should account for the ethical aspects.
You need good design for evaluation.

The AI Hype


Oh, so was this just an overestimation of AI?
No claims should be made that MT systems match humans if they only use AEMs, for example. (Moorkens, 2022)

The AI Hype

Neural Systems

- Often equated with artificial intelligence (AI)
- Significant progress in system capabilities

(Koehn, 2020, p. 33)

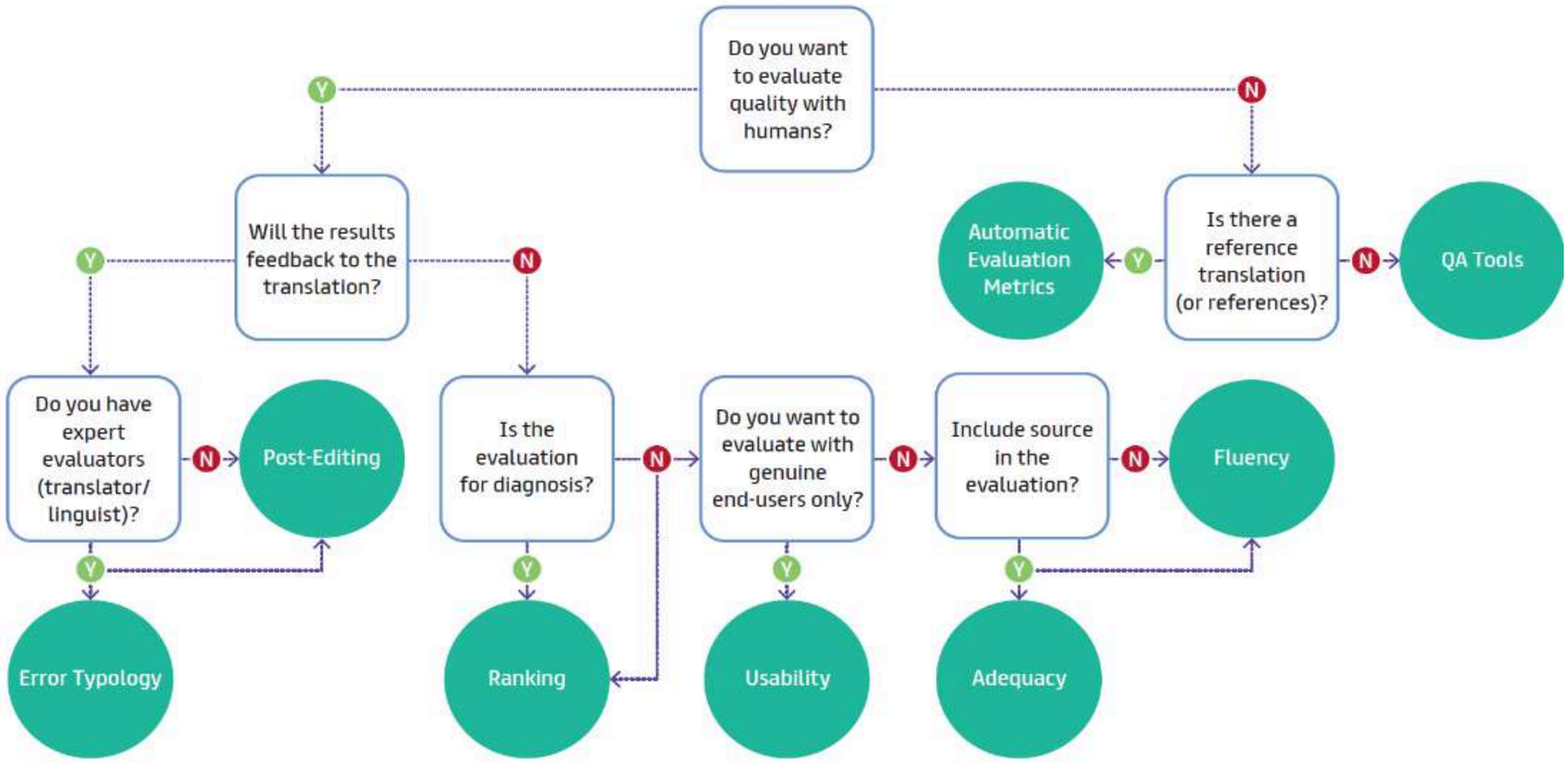
Human Expertise

- Both assessment and editing of NMT ouput is still required in the same way

(Ragni and Vieira, 2022, p. 148)

Remember! Critical tasks should not be entrusted solely to machines (Floridi et al., 2021)

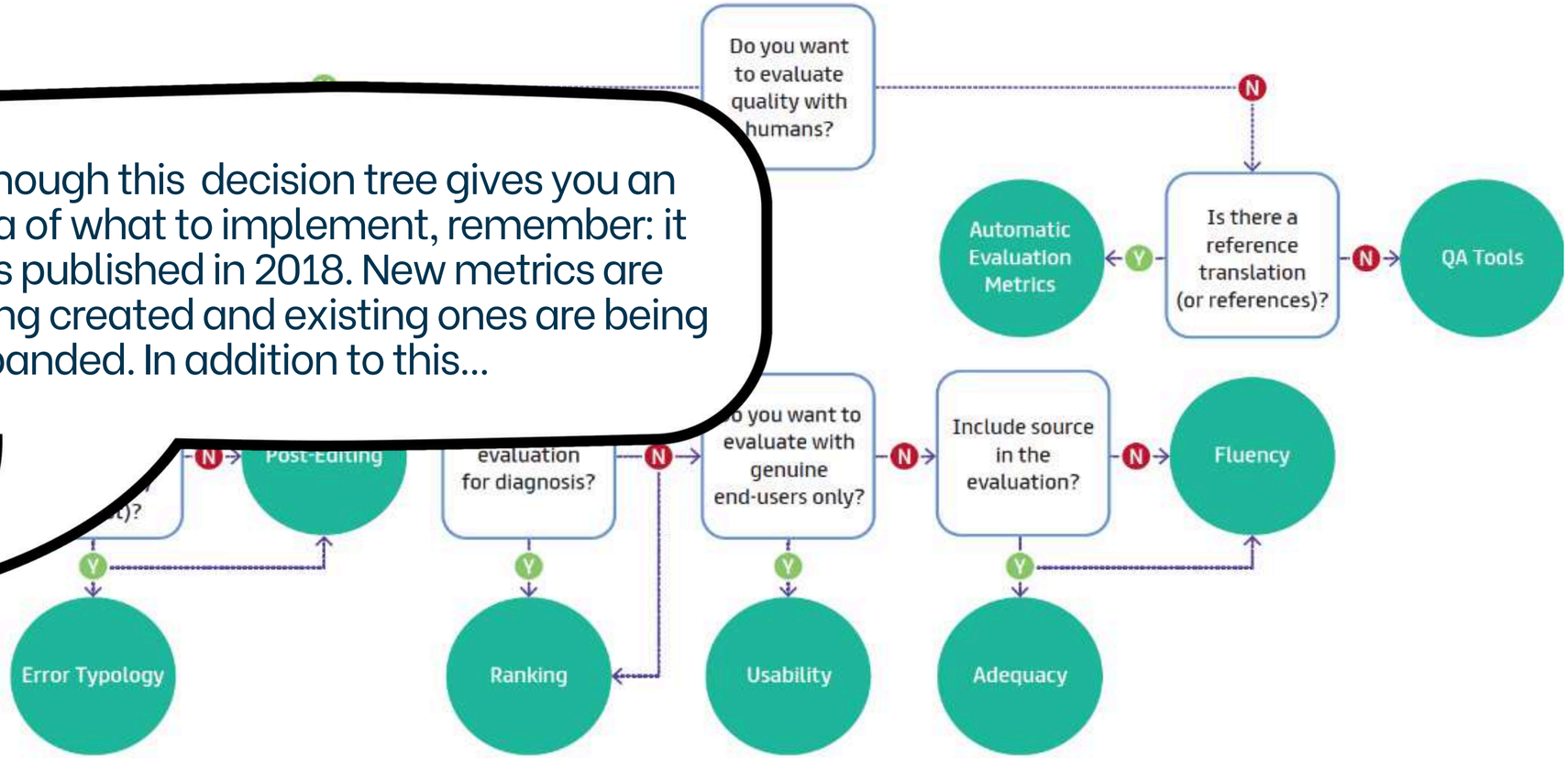
Evaluation – Where do I start?



Doherty et al. (2018)

Evaluation – Where do I start?

Although this decision tree gives you an idea of what to implement, remember: it was published in 2018. New metrics are being created and existing ones are being expanded. In addition to this...



Doherty et al. (2018)

Evaluation – Where do I start?

Perishability

"Texts are considered perishable if they are for immediate consumption with little or no purpose thereafter" (p. 100)

"Non-perishable texts (literary works and marketing copy) are typically carefully crafted so as to possess aesthetic value and/or to clearly convey important, often durable, messages. (p. 100)

Doherty et al. (2018)

Metrics

There is a range of measures performed by humans that have been operationalised in the field of quality evaluation. (Castilho et al., 2018, 17)

- The most common and employed is a combination of Adequacy/Fluency.

There are other measures such as:

- Readability
- Comprehensibility
- Usability
- Acceptability
- Ranking

Castilho et al. (2018)

Metrics – Adequacy/Fluency

Adequacy

- Also called “accuracy” or “fidelity” in certain studies.
- It is the “extent to which the translation transfers the **meaning of the source-language** unit into the target”

Fluency

- Also called “intelligibility” (which may have other meanings in specific contexts)
- It is the “the extent to which the translation follows the **rules and norms of the target-language** (regardless of the source or input text).”

- Adequacy and Fluency typically are used together!!

Castilho et al. (2018)

Metrics – Adequacy/Fluency

- Typically assessed using ordinal scales through Likert scales, for example:

Adequacy

"How much of the meaning in the source appears in the target?"

1. None of it
2. Little of it.
3. Most of it.
4. All of it.

Fluency

"How fluent is the translation?"

1. No fluency.
2. Little fluency.
3. Near native.
4. Native

- Adequacy requires some degree of bilingual proficiency. Fluency requires only proficiency in the target language.
- Choose your **professional evaluators** wisely!

Castilho et al. (2018)

Metrics – Error Classification

- Predominantly the evaluation model of the industry.

- Translated texts or samples of translated texts have the errors identified, counted and have a weight according to how severe the error is.
- A common assessment approach in academic translator training programmes.

- It began with the Localisation Industry Standards Associsation (LISA) QA Model, in the late 1990s.

Castilho et al. (2018)

Metrics – Error Classification

- The goal is usually to identify and classify errors in the output.

Different taxonomies have been proposed:

- Llitjós et al. (2005)
- Vilar et al. (2006)
- Federico et al. (2014)
- Costa et al. (2015)
- **DQF - TAUS**
- **MQM-QT21**

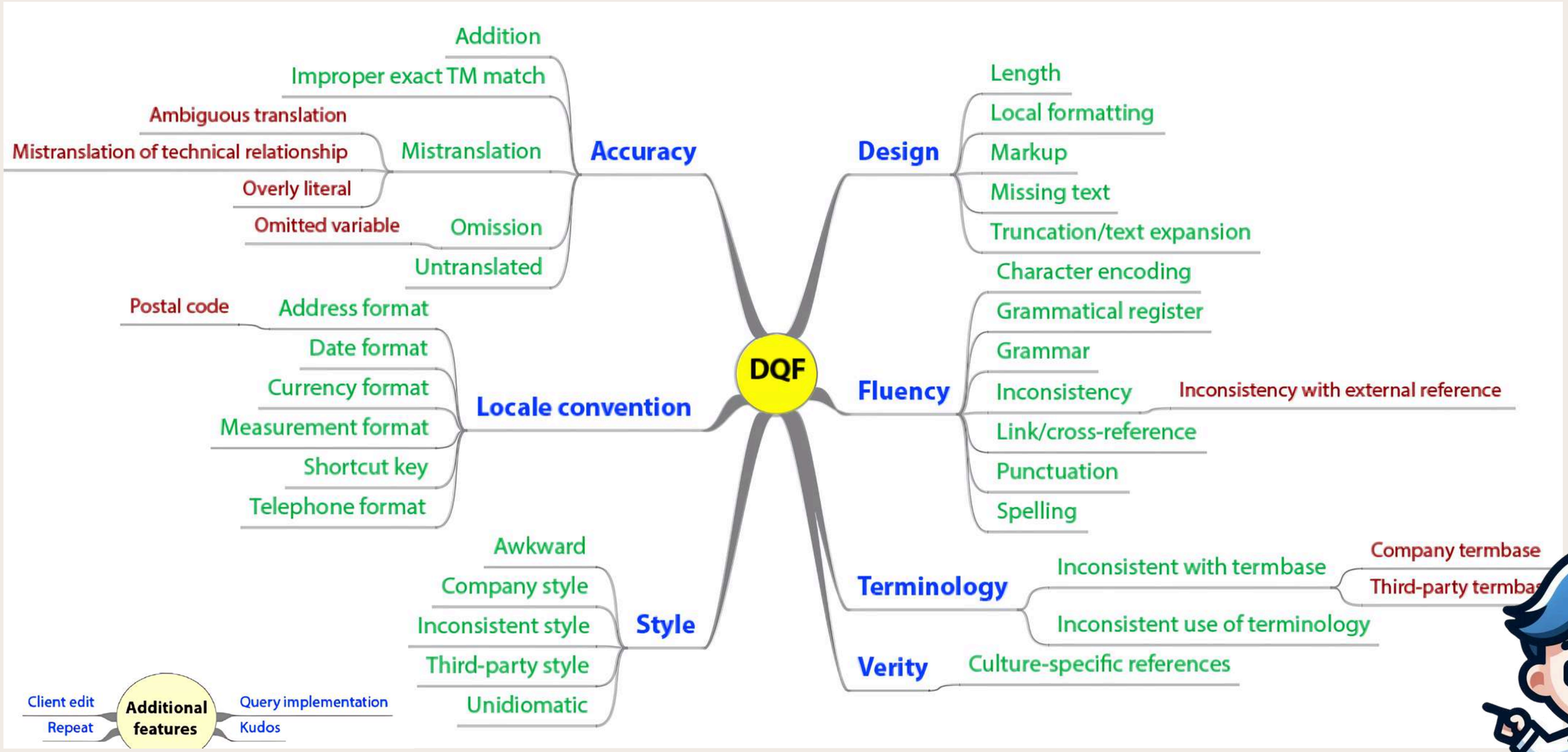
- You may find the criteria here: https://themqm.org/error-types-2/typology/

Metrics – Error Classification

- The **Translation Automation User Society (TAUS)** is responsible for the Dynamic Quality Framework (DQF).
- The initiative **QTLaunchPad (and later QT21)** launched the Multidimensiona quality Metrics (MQM) framework.
- Most recent large-scale initiatives to standardise this type of assessment.

Castilho et al. (2018)

Metrics – Error Classification



Metrics - Acceptability

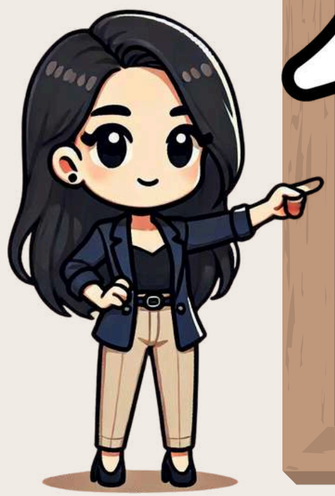
- Used in various fields, such as linguistics, translation and human-computer interaction (HCI).
- Definitions may vary according to the field and author.

- It is the extent that users will accept, reject or tolerate when reading a MT output.
- Other factors such as usability, satisfaction and quality may impact this metric.

Castilho et al. (2018)

Metrics - Acceptability

Castilho (2016) found that users will find translations acceptable if they can use it to perform tasks, even if the translation has flaws.



- ...ch as linguistics, translation ...raction (HCI).
- ...t is the extent that users will accept, reject or tolerate when reading a MT output.
- Other factors such as usability, satisfaction and quality may impact this metric.

Castilho et al. (2018)

Metrics - Ranking

- Typically used in research contexts.
- It compares output from different MT systems.
- Typically evaluators are asked to rank given sentences in the target-language.
- Ranking may use criteria such as Adequacy/Fluency or may allow a subjective choice of what seems to be best according to the opinion of evaluators.

- The MT outputs (two or more) are listed in a **random scrambled order**, they have to be **unpredictable** and they must be **anonymous**.

Castilho et al. (2018)

Metrics - Ranking - Exercise

- **Source:** Essas temporadas são as versões MsD mais antigas, e as temporadas 2-5 NÃO foram anunciadas ou lançadas pela Shout Factory

- **Option 1:** These periods are the earliest MsD versions, and the periods 2-5 were NOT notified or issued by the Shout Factory
- **Option 2:** These series are the oldest MsD editions, and series 2-5 were NOT announced or distributed by Shout Factory.
- **Option 3:** These seasons are the oldest MsD versions, and seasons 2-5 have NOT been announced or released by Shout Factory

Metrics - Usability

- Tested with real end-users, on how they use the product or service in which there is translated content.
- The **ISO/TR 16982** definition- “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (International Organization for Standardisation, 2002).

- It is:

Subjective

- Opinions
- Satisfaction
- Recommendation

AND

Objective

- Time
- Browsing behaviour

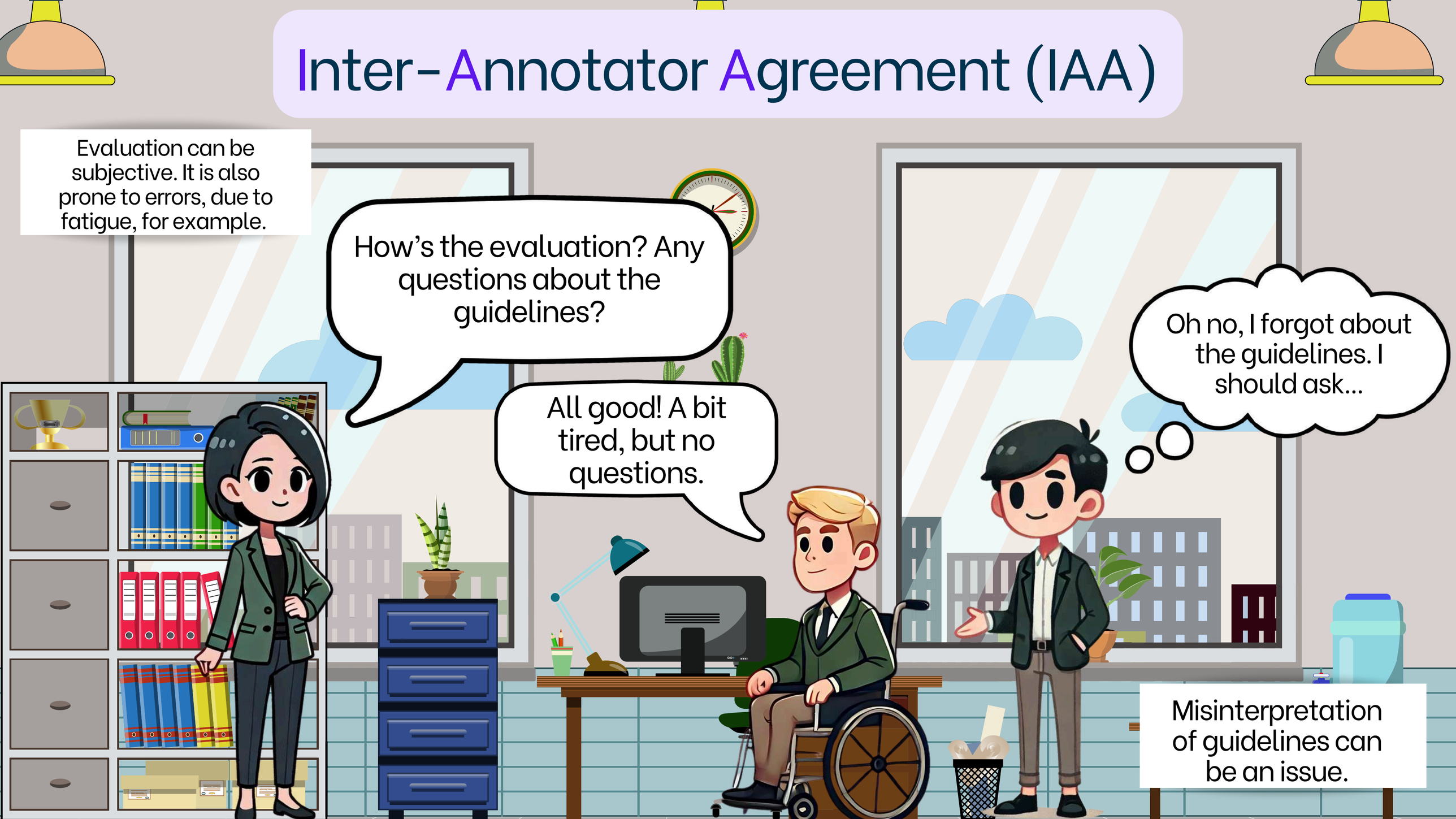
Castilho et al. (2018)

Inter-Annotator Agreement (IAA)

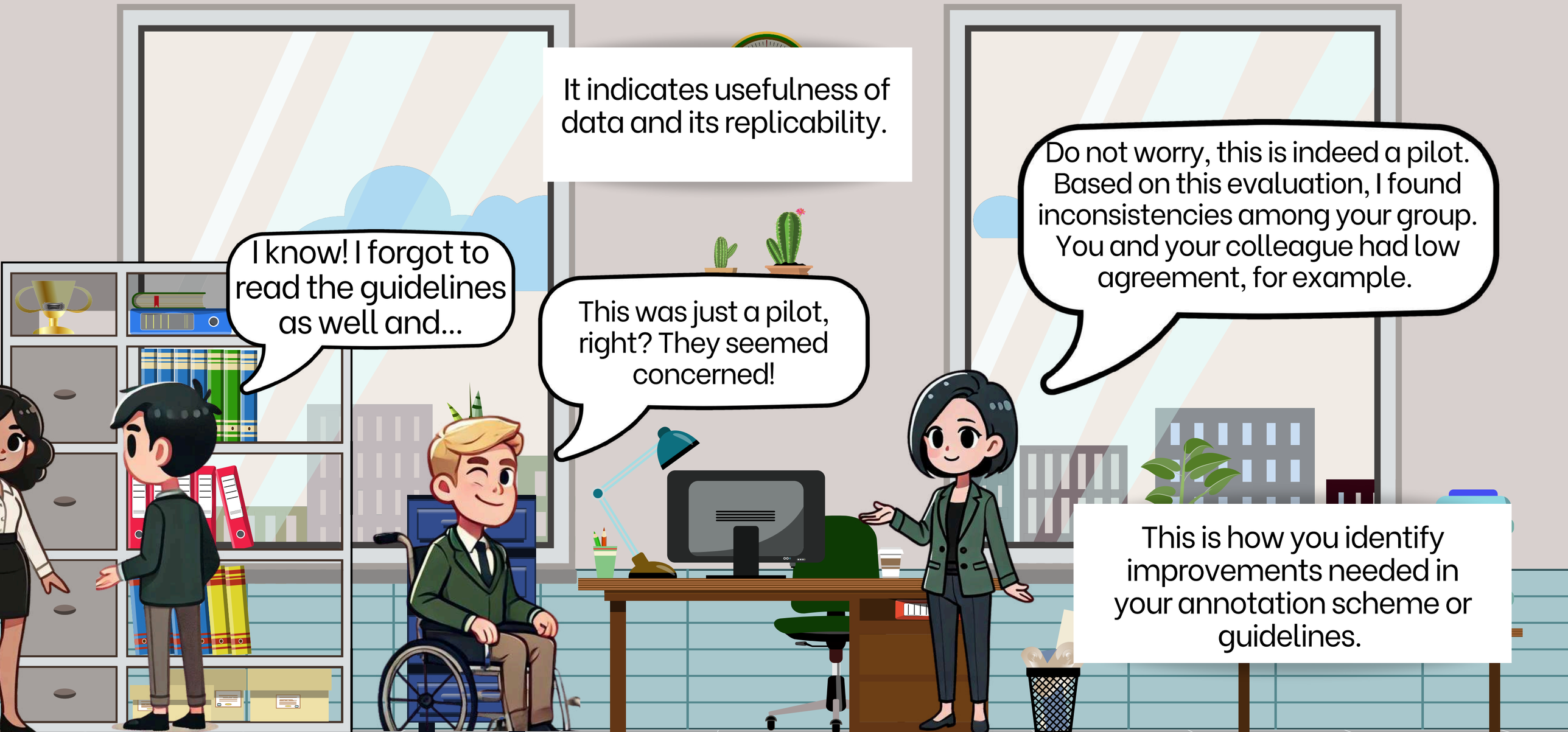
- It is a measure of how multiple annotators can make the same decision.
- It shows how clear the annotation guidelines are.
- How uniformly annotators understood guidelines.
- How reproducivble the annotation task is.

It is vital for **validation** and **reproducibility**!!

Inter-Annotator Agreement (IAA)



Inter-Annotator Agreement (IAA)



Inter-Annotator Agreement (IAA)

The most used coefficients are:

- Cohen's Kappa (weighted and non-weighted)
- Fleiss' Kappa

Coeficients	Chance Correction	Weighted	# Raters	Measurement
Inter-rater reliability (IRR)	no	no	any	percentage
Cohen's Kappa	yes	no	2	interval 0-1
Weighted Cohen's Kappa	yes	yes	2	interval 0-1
Fleiss' Kappa (version of Scott's)	yes	no	any	interval 0-1
Krippendorff's Alpha	yes	yes	any	interval 0-1

Castilho (2021)

They are common statistics in the field of computational linguistics. It helps to determine whether the assessments capture observable reality to a certain degree. (Artstein and Poesio, 2008)

Inter-Annotator Agreement (IAA)

Coeficients	Chance Correction	Weighted	# Raters	Measurement
Inter-rater reliability (IRR)	no	no	any	percentage
Cohen's Kappa	yes	no	2	interval 0-1
Weighted Cohen's Kappa	yes	yes	2	interval 0-1
Fleiss' Kappa (version of Scott's)	yes	no	any	interval 0-1
Krippendorff's Alpha	yes	yes	any	interval 0-1

- **Cohen's Kappa** (Cohen, 1960) can be both non-weighted and weighted. It measures the agreement between only two raters.
- **Fleiss' Kappa** (Fleiss, 1971) accounts for more than two raters.
- **Krippendoff's Alpha reliability** (Krippendoff, 2011) applies to multiple coders, allows for different magnitudes of disagreement).
- **Inter-Rater Agreement (IRR)** can also be calculated, of the number of agreements divided by the total number of assessments.

Inter-Annotator Agreement (IAA)

How to interpret the agreement?
Example

- ≤ 0 indicates no agreement.
- 0.01–0.20 as none to slight agreement
- 0.21–0.40 as fair agreement
- 0.41–0.60 as moderate agreement
- 0.61–0.80 as substantial agreement
- 0.81–1.00 as almost perfect agreement

Inter-Annotator Agreement (IAA)

How to interpret the agreement? Example

- ≤ 0 indicates no
- 0.01–0.20 as no
- 0.21–0.40 as fa
- 0.41–0.60 as mo
- 0.61–0.80 as substantial agreeme
- 0.81–1.00 as almost perfect agreeme

There is not a specific consensus. This is one interpretation, as they may vary depending on the community. (Artstein and Poesio, 2008)

In this lecture you were able to...

Understand why we perform evaluation, the strengths and weaknesses of human evaluation and automatic evaluation.

Understand that different systems have different purposes, and therefore, the type of evaluation and metrics must be adjusted accordingly.

Understand how different metrics can be employed in the evaluation of MT systems.

Thank you! See you next class. Questions?

Send an e-mail to joo.cavalheirocamargo2@mail.dcu.ie