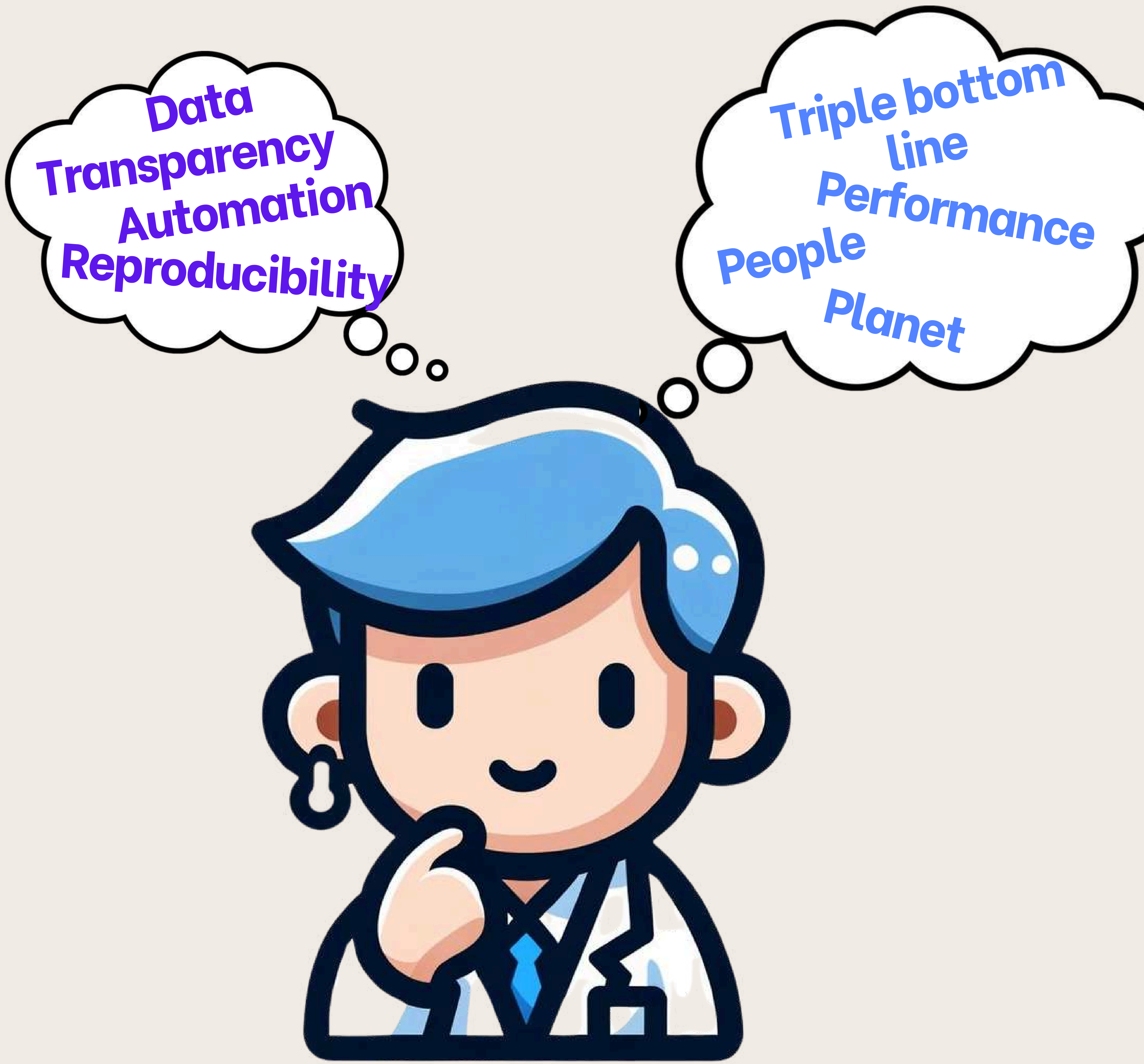
DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Machine Translation Quality Assessment Lesson 4 - Ethics in Evaluation



M.Ed. João Lucas Cavalheiro Camargo

Learning Outcomes

LO4 - Design replicable evaluation of MT systems, cognisant of the diverse evaluation approaches and types of evaluators in the process.

LO5 - Report results from the evaluation of an MT system addressing the context, the type of evaluators and the use case of the MT system.

Structure

1 - Recap

2 - Going over the hype

3 - To what degree should we automate?

4 – Impacts on people: Ethical considerations on translators

5 - Impacts on research: Ethical considerations for MT research

6 - Towards a Triple Bottom Line (TBL)

7 - TBL - People

8 - TBL – Planet

9 – TBL – Performance

Going over the hype


Going over the hype
We have progressed a lot in MT, so is translation a solved problem?
We've come a long way, but it's far from solved. The evaluation issues we've covered show that MT still struggles in many areas. And there's growing research that highlights these challenges.

Is MT really solved?

Productivity in the Post-editing of Neural Machine Translation: A Mixed Methods Analysis of Speed and Edits at Toppan Digital Language

Terribile (2024)

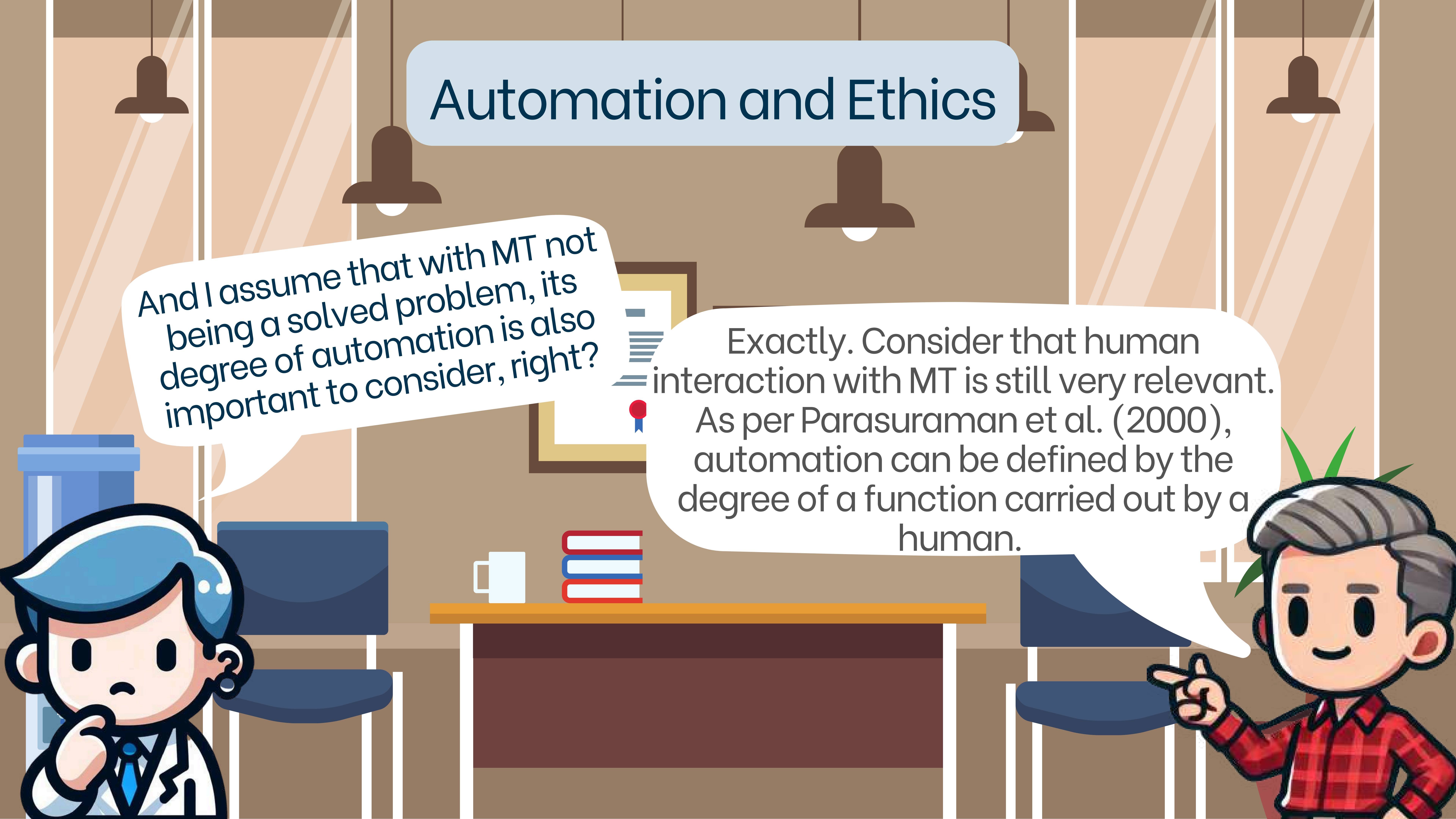
- Investigated over two and a half years over ninety million words post-edited between major European languages.
- Terribile reports that over 40% of all edits involved a significant alteration
- NMT is usually unable to retrieve information that is left implicit in the source (p. 231)

Is MT really solved?

Automating Translation

Moorkens et al. (2024)

- ✅ They report that projects such as European Language Equality show that tools and data are widely available only for English, and to a lesser extent for other languages such as French and Spanish.
- ✅ We still need to understand better how and why AI models work
- ✅ Explainable AI (xAI) should be a goal.


Automation and Ethics
And I assume that with MT not being a solved problem, its degree of automation is also important to consider, right?
Exactly. Consider that human interaction with MT is still very relevant. As per Parasuraman et al. (2000), automation can be defined by the degree of a function carried out by a human.

Automation and Ethics

A Model for Types and Levels of Human Interaction with Automation

Parasuraman et al. (2000)

- Defines that automation substitutes human involvement in a task, either completely or to some degree (p. 287)
- They propose four stages of human information processing: Sensory processing, perception/working memory, decision making, response selection
- Levels of automation can help make informed decisions about development, evaluation and use.

Automation and Ethics

A Model for Types and Levels of Human Interaction with Automation

Parasuraman et al. (2020)

HIGH

10. The computer decides everything, acts autonomously, ignoring the human.
9. informs the human only if it, the computer, decides to
8. informs the human only if asked, or
7. executes automatically, then necessarily informs the human, and
6. allows the human a restricted time to veto before automatic execution, or
5. executes that suggestion if the human approves, or
4. suggests one alternative
3. narrows the selection down to a few, or
2. The computer offers a complete set of decision/action alternatives,
1. The computer offers no assistance: human must take all decisions and actions

LOW

Automation and Ethics

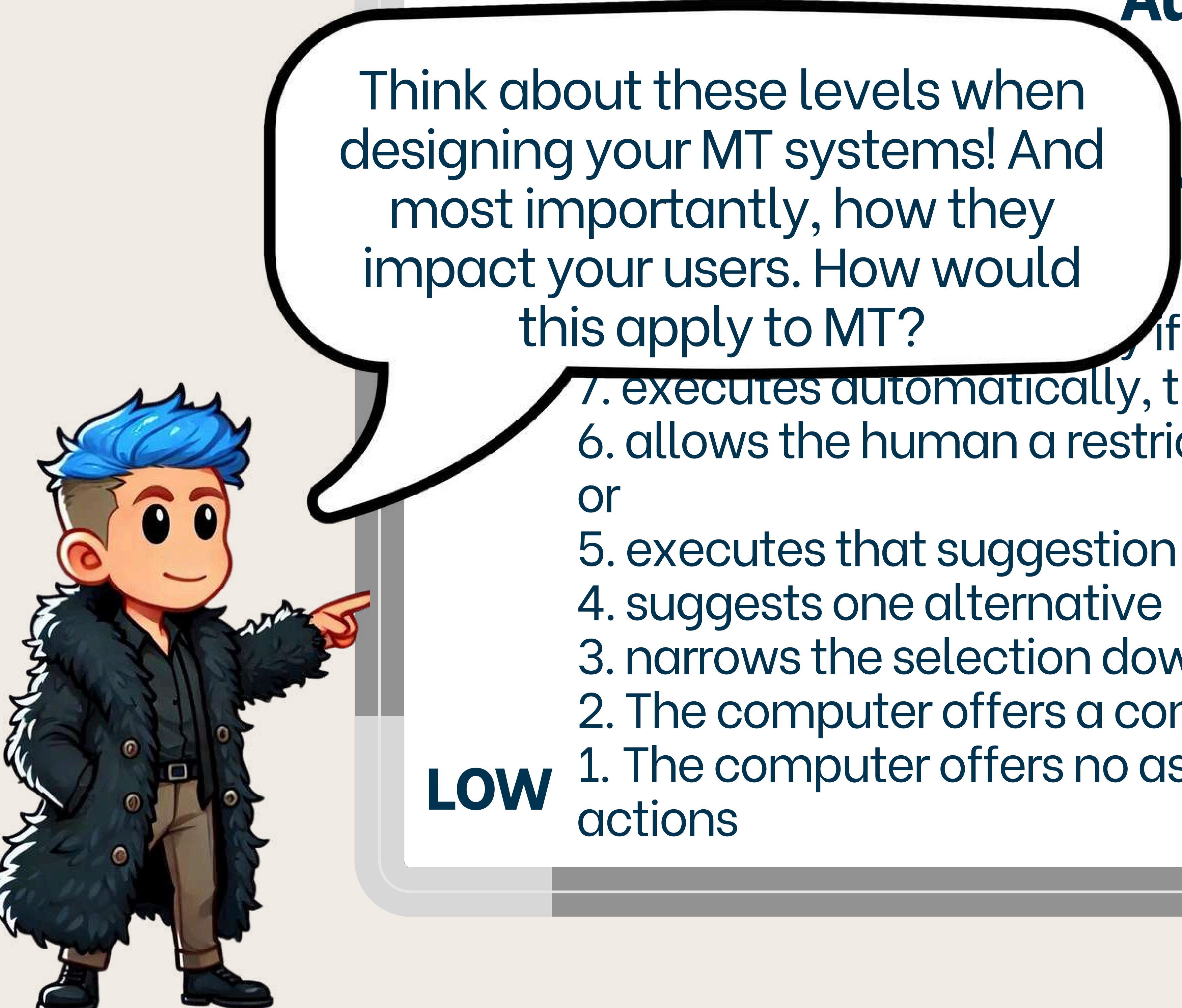
Level of Automation	Description
Level 1	**Manual control** Computer offers no assistance
Level 2	**Decision proposal stage** Computer suggests decisions, operator selects and executes
Level 3	**Human decision select stage** Human selects a decision, computer executes
Level 4	**Computer decision select stage** Computer selects decision, executes with human approval
Level 5	**Computer execution and human info** Computer executes, informs human
Level 6	**Computer execution and on-call human info** Executes, informs human only if asked
Level 7	**Computer execution and voluntary info** Executes, informs only if needed
Level 8	**Autonomous control** Computer does everything, informs only in case of error

Vagia et al. (2016)

Automation and Ethics

A Model for Types and Levels of Human Interaction with Automation

Parasuraman et al. (2020)



…verything, acts autonomously, ignoring the
… it, the computer, decides to
… if asked, or
7. executes automatically, then necessarily informs the human, and
6. allows the human a restricted time to veto before automatic execution, or
5. executes that suggestion if the human approves, or
4. suggests one alternative
3. narrows the selection down to a few, or
2. The computer offers a complete set of decision/action alternatives,

LOW 1. The computer offers no assistance: human must take all decisions and actions

Automation and Ethics

HT = Human translation
CAT = Computer Assisted Translation
TM = Translation Memory

There are different levels of automation in translation and MT. How would you classify them?

Human Translation

HT-aided by CAT tools

Light Post-Editing

HT-aided by TMs

CAT + MT integration

Raw MT

Adaptive MT

Heavy Post-editing

Automation and Ethics

HT = Human translation
CAT = Computer Assisted Translation
TM = Translation Memory

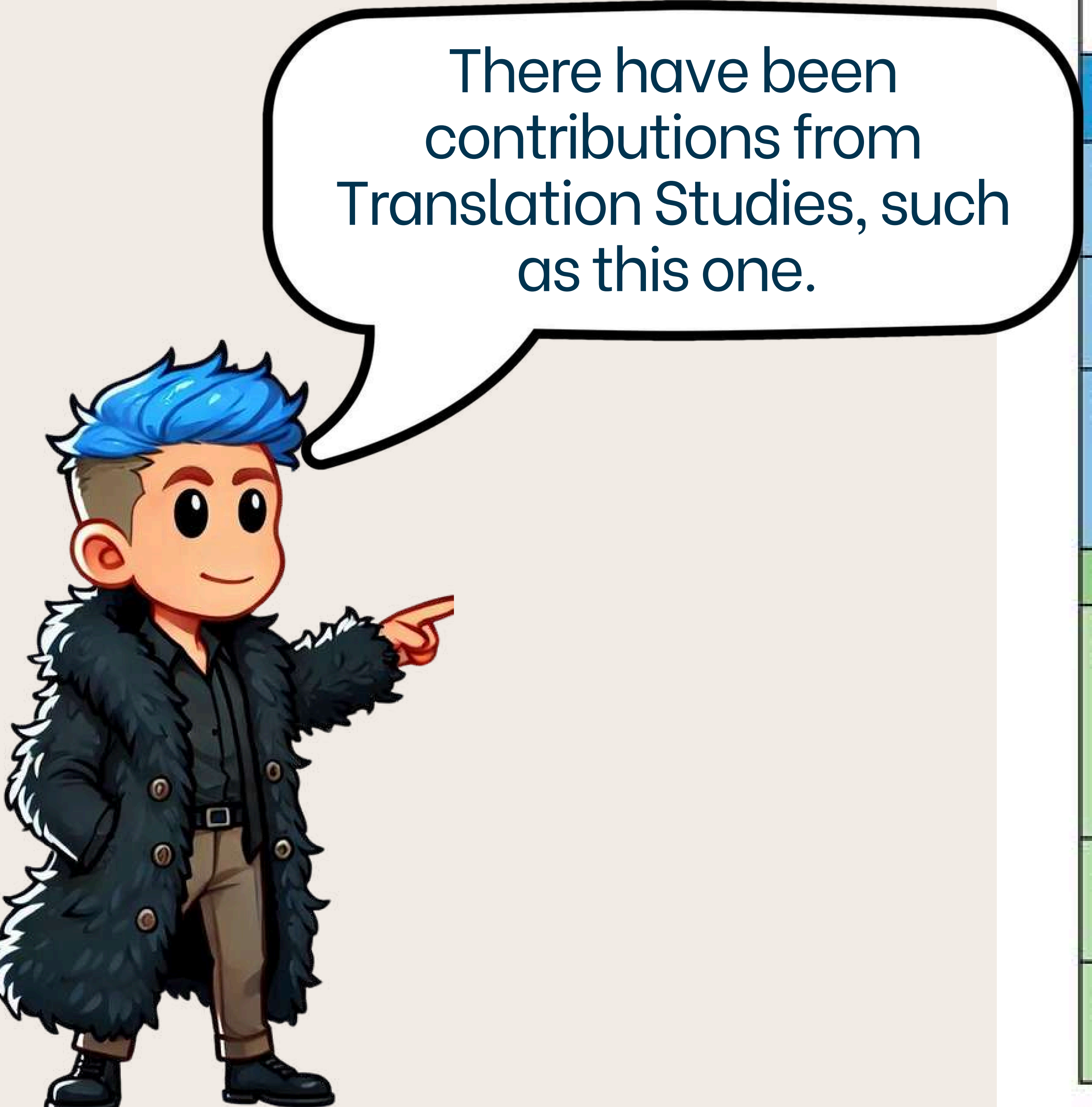
There is no consensus. But imagine we combined Vagia et al. (2016)'s taxonomy with translation

Level of Automation	Description	Type of Translation
Level 1	**Manual control** Computer offers no assistance	a) Human Translation
Level 2	**Decision proposal stage** Computer suggests decisions, operator selects and executes	a) Human Translation aided by TM b) Human Translation aided by CAT tools
Level 3	**Human decision select stage** Human selects a decision, computer executes	a) Human Translation aided by TM b) Human Translation aided by CAT tools c) Post-Editing of Adaptive Machine Translation
Level 4	**Computer decision select stage** Computer selects decision, executes with human approval	a) Human Translation aided by Translation Memories b) Human Translation aided by Computer Assisted Tools c) Post-Editing of Adaptive Machine Translation d) Light Post-Editing of Machine Translation e) Heavy Post-Editing of Machine Translation f) Light/Heavy Post-Editing of MT + CAT tool
Level 5	**Computer execution and human info** Computer executes, informs human	
Level 6	**Computer execution and on-call human info** Executes, informs human only if asked	
Level 7	**Computer execution and voluntary info** Executes, informs only if needed	
Level 8	**Autonomous control** Computer does everything, informs only in case of error	a) Raw Machine Translation

Automation and Ethics

Christensen et al. (2021)

Level	Name	Dynamic Translation Task (DTT): Control of source text analysis and target text production	Dynamic Translation Task (DTT): Error and inadequacy detection and response	DTT Fallback	Operational Design Domain (ODD)
Translator performs all or part of the DTT					
0	**No TA**	Translator	Translator	Translator	n/a
1	**Translator Assistance**	Translator and System	Translator	Translator	Limited
2	**Partial TA**	System	Translator	Translator	Limited
"Automated translation system" (ATS "system") performs the entire DTT					
3	**Conditional TA**	System	System	Fallback-ready user (becomes the translator during fallback)	Limited
4	**High TA**	System	System	System	Limited
5	**Full TA**	System	System	System	Unlimited



Automation and Ethics

HT = Human translation
CAT = Computer Assisted Translation
TM = Translation Memory

Level of Automation	Description	Type of Translation
Level 1	The computer offers no assistance: human must take all decisions and actions	a) Human Translators using Pen-and-paper or Mechanical Typewriters (Rare/Unlikely Scenario)
Level 2	The computer offers a complete set of decision/action alternatives, or	a) Human Translation aided by CAT tools - Spelling and Grammar checking
Level 3	The computer narrows the selection down to a few, or	a) Human Translation aided by TM and terminology suggestions
Level 4	The computer suggests one alternative	
Level 5	The computer executes that suggestion if the human approves, or	a) Edition or Approval of MT
Level 6	The computer allows the human a restricted time to veto before automatic execution, or	a) The extent of MT that can be approved or edited
Level 7	The computer executes automatically, then necessarily informs the human, and	
Level 8	The computer informs the human only if asked, or	
Level 9	The computer informs the human only if it, the computer, decides to	
Level 10	The computer decides everything, acts autonomously, ignoring the human.	a) Raw MT

Automation and Ethics

Artificial Intelligence, automation and the language industry

Moorkens and Guerberof Arenas (2024)

- The integration of translation technologies and MT in platforms varies, so it is impossible to measure the extent of how much MT is used in human translation.
- The gathering of translation data and translator activity data and other types of data may affect how translators get offered jobs.
- In the audiovisual translation industry, automation has led to less payment (p. 81)

Automation and its impact on translators

Taking a first step: what is Ethics?

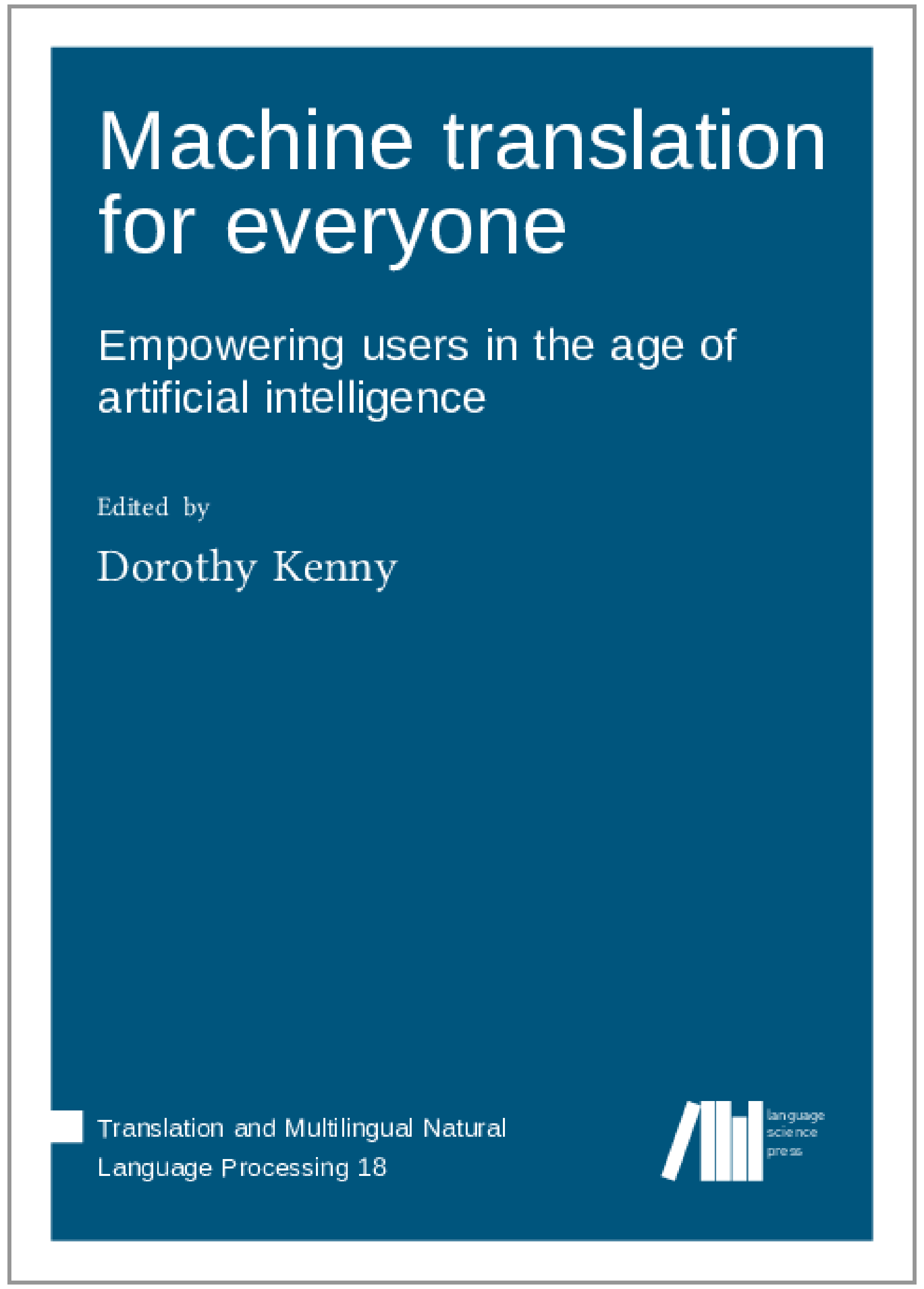
Moorkens (2022)

- Ethics is the field that examines morality, good and evil, right and wrong, etc.
- Philosophers and ethicists have worked on different courses of action, based on what is right or moral, based on outcomes that would benefit the majority.
- Applied ethics is the field that aims to address specific problems. Normative ethics provide the rationale for the application of ethical behaviour or solutions

Automation and its impact on translators

Data and Ownership

- Optional resource: Read p. 122–123 for a case study on data ownerhip



Moorkens (2022)

Automation and its impact on translators

Data and Ownership

Moorkens (2022)

- Human data is VALUABLE! MT training data are stored as parallel (or aligned) bilingual segments of text, translated by humans, stored in databases called translation memories.
- If translation databases are being used, who has ownership rights? The translator? Is that being respected?
- In practice, translation memories are sent to clients.

Automation and its impact on translators

Data and Ownership - Use

Moorkens (2022)

- Depends on the jurisdiction. In some, whoever pays owns the translation, while in others, ownership may be transferred.
- Data may have metadata attached to it - name/IDs, date and time of creation, language codes, software used, a project ID.
- When translation platforms are used, other data can be collected, such as activity data of translators with detailed timings, editing actions and even the records of keystrokes.

Automation and its impact on translators

Data and Ownership - Distribution

Moorkens (2022)

- Agreements between companies and organisations may lead to the distribution of data.
- Data can be bought, sold or donated for research purposes.
- With regulations to be concerned, personal data has restrictions in its distribution.

Automation and its impact on evaluation

Ethics in MT evaluation

Moorkens (2022)

- There are also ethical issues in MT evaluation.
- Most of the output of MT systems is evaluated using automatic methods during training for quick, easy and cost-effective measures.
- In shared tasks, development teams use either automatic or crowd evaluation typically.
- Sometimes automatic evaluation with segment-level crowd rating can be reported to reach parity with human translation quality.

Automation and its impact on evaluation

Ethics in MT evaluation

Moorkens (2022)

- Language is important! Reporting the capabilities of our systems must match how your evaluation was performed.
- If the capability of MT systems are overestimated, that might lead to media reproducing that attitude.
- Crowd workers will never match the same level of evaluation as expert evaluators.
- Crowd workers can suffer poor working conditions: pay, labour conditions, used as research participants without ethical review.

Automation and its impact on evaluation

Working Conditions of Translators

Moorkens (2022)

- Translation is a HIGHLY SKILLED TASK. But automation has impacted translators' jobs.
- Changes have included: economic returns, work organisation and skills management.
- Translators are largely freelance, which led to dependency on project-by-project conditions.
- These conditions led translators to have little say in processes that are changed unilaterally by agencies and employers.

Automation and its impact on evaluation

Working Conditions of Translators

Moorkens (2022)

- It also has impacted how translation has been performed – translators' work may include quality checks, annotation or correction of repetitive errors from MT output.

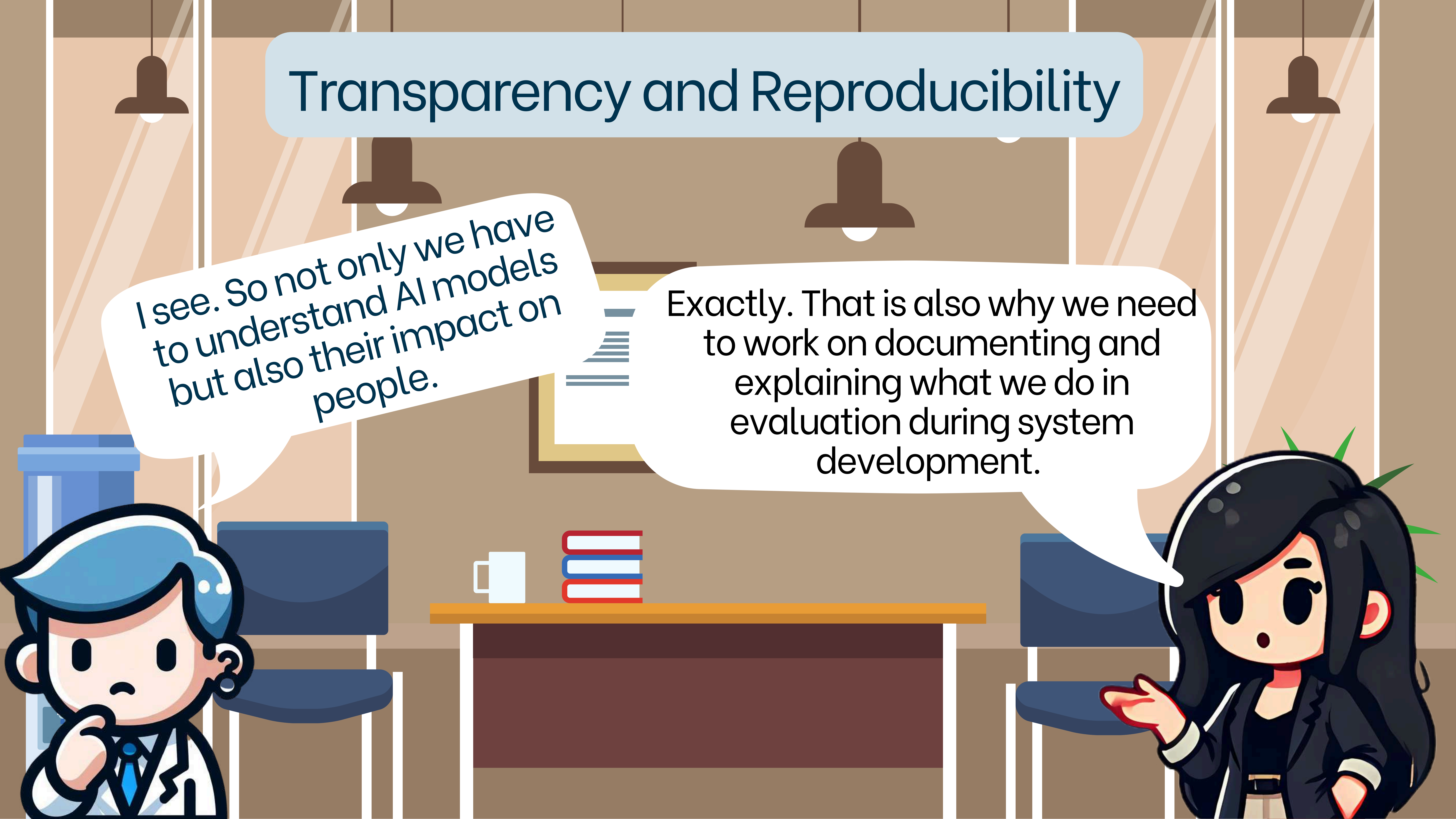
- Satisfaction impacted the profession as well – with some translators enjoying post-editing, and others disliking it, due to reasons such as discounts in payment.

Automation and its impact on evaluation

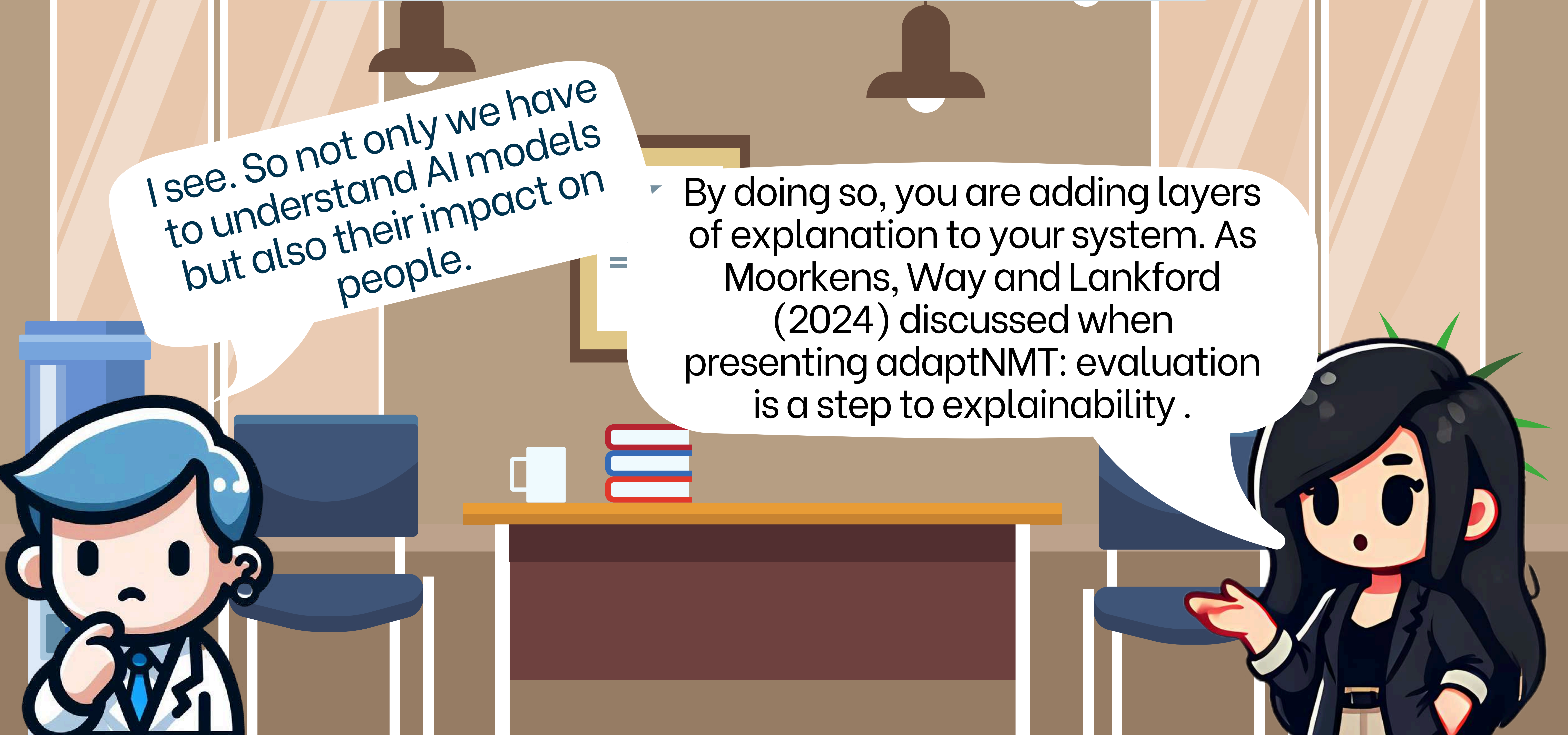
Working Conditions of Translators

Moorkens (2022)

- ✅ With AI being adopted, more AI related services have been offered by companies. For example, data generation, annotation, validation, chatbot text generation, testing, engineering and synthetic data creation.


Transparency and Reproducibility
I see. So not only we have to understand AI models but also their impact on people.
Exactly. That is also why we need to work on documenting and explaining what we do in evaluation during system development.

Transparency and Reproducibility



Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- The researchers manually annotated MT evaluation papers published from 2010 to 2020 at ACL conferences.
- They annotated the automatic metrics used.
- They annotated whether human evaluation had been conducted, if yes or no.
- They annotated whether any type of statistical significance testing was performed.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- They annotated whether papers made comparison of automatic scores by copying them from previous work.
- They annotated whether SacreBLEU was used or not.
- They annotated whether previous work had been reproduced or copied. (e.g. if the authors used the same pre-processed training, validation and testing data).

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Main issue found 1: Majority of the papers used BLEU.
- Relying solely on BLEU scores without statistical significance testing nor human evaluation can lead to the wrong conclusions in evaluation!
- The authors recommend other metrics to better correlate with human judgments.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Main issue found 2: No Statistical Significance Testing
- We use statistical significance testing to ensure that results of experiments do not happen by chance.
- For each year verified by the authors, never more than 65% of the publications performed statistical significance testing.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Main issue found 3: Results are Copied
- When comparing MT systems with previous work, sometimes the paper copies the scores reported on papers published.
- Researchers found that most papers do not find enough information to enable papers to be compared (Post 2018).
- SacreBLEU should help with standardisation.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Main issue found 3: Results are Copied
- Since BLEU requires several parameters and is dependent on pre-processing of the MT output and reference translation, it is difficult to replicate results.
- Depending on the tokenisation of your MT output, it can vastly affect BLEU scores!
- If using SacreBLEU, make sure to include the signature so scores can be reproduced.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Main issue found 4: Data Approximation
- Pre-processing of datasets matter - during the training, the tuning and the evaluation.
- Differences in tokenisation, casing and length filtering impact the scores.
- Because of these differences, papers could be making comparisons and conclusions on a flawed basis.

Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Guidelines
- Do NOT rely exclusively on BLEU.
- Perform statistical significance testing on automatic metric scores. Ensure the difference and amplitude is not by chance.
- If comparing score, make sure they are being computed on the same way.

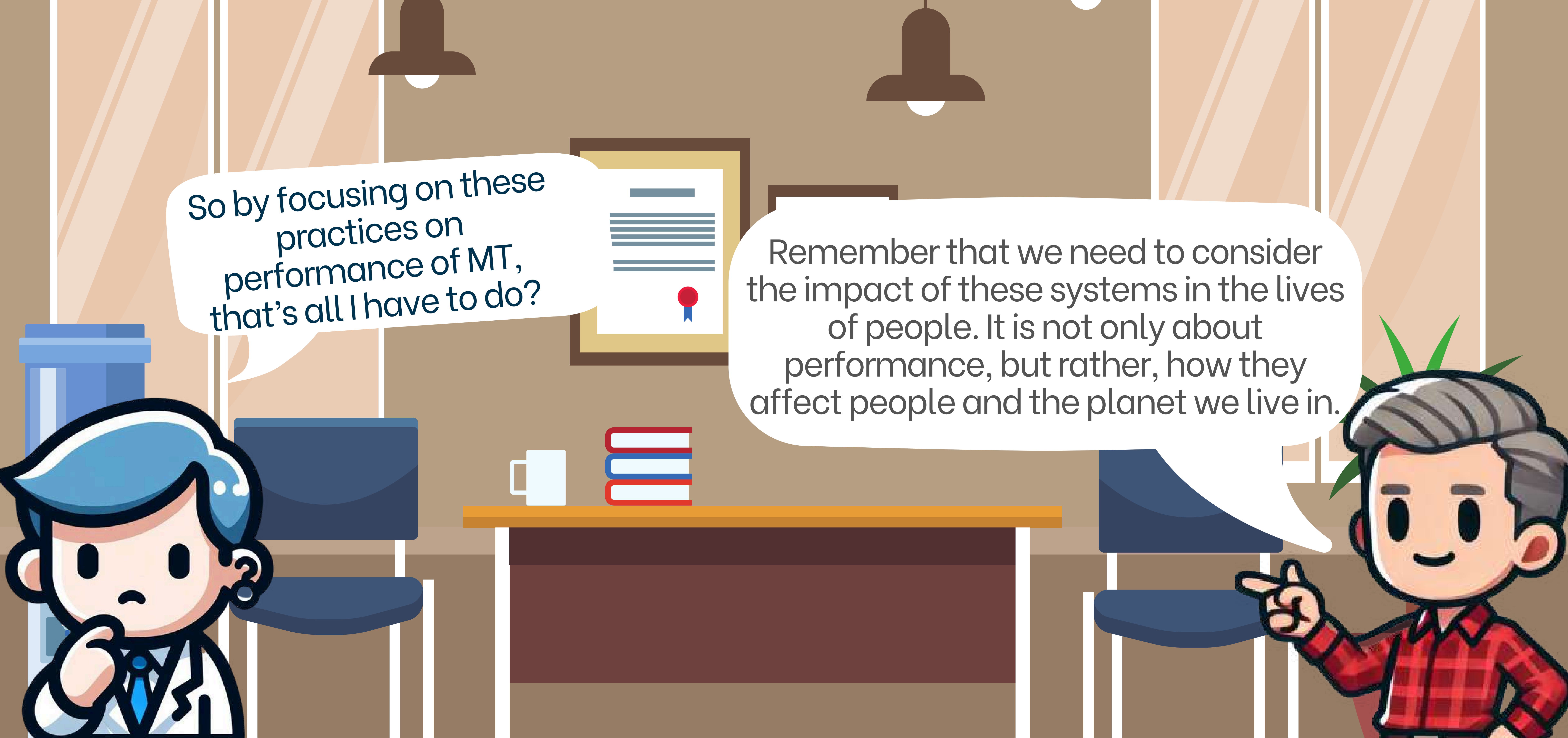
Transparency and Reproducibility

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Marie et al. (2021)

- Guidelines
- When comparing MT systems through metric scores to demonstrate the superiority of a method or algorithm, only do that if the systems have been trained, validated and tested with exactly the same pre-processed data.
- That does not applied when the proposed method or algorithm is dependent on a particular dataset or pre-processing.

What's next?


So by focusing on these practices on performance of MT, that's all I have to do?
Remember that we need to consider the impact of these systems in the lives of people. It is not only about performance, but rather, how they affect people and the planet we live in.

Towards a Triple Bottom Line (TBL)

Translation, technology and Climate change

Cronin (2019)

- Desktops, laptops, and data centers, are significant contributors to carbon emissions due to their increasing energy consumption and reliance on fossil fuels.

- The rapid expansion of technology accelerates energy demand, creating an upward spiral of environmental impact and contributing to climate change.

Towards a Triple Bottom Line (TBL)

Translation, technology and Climate change

Cronin (2019)

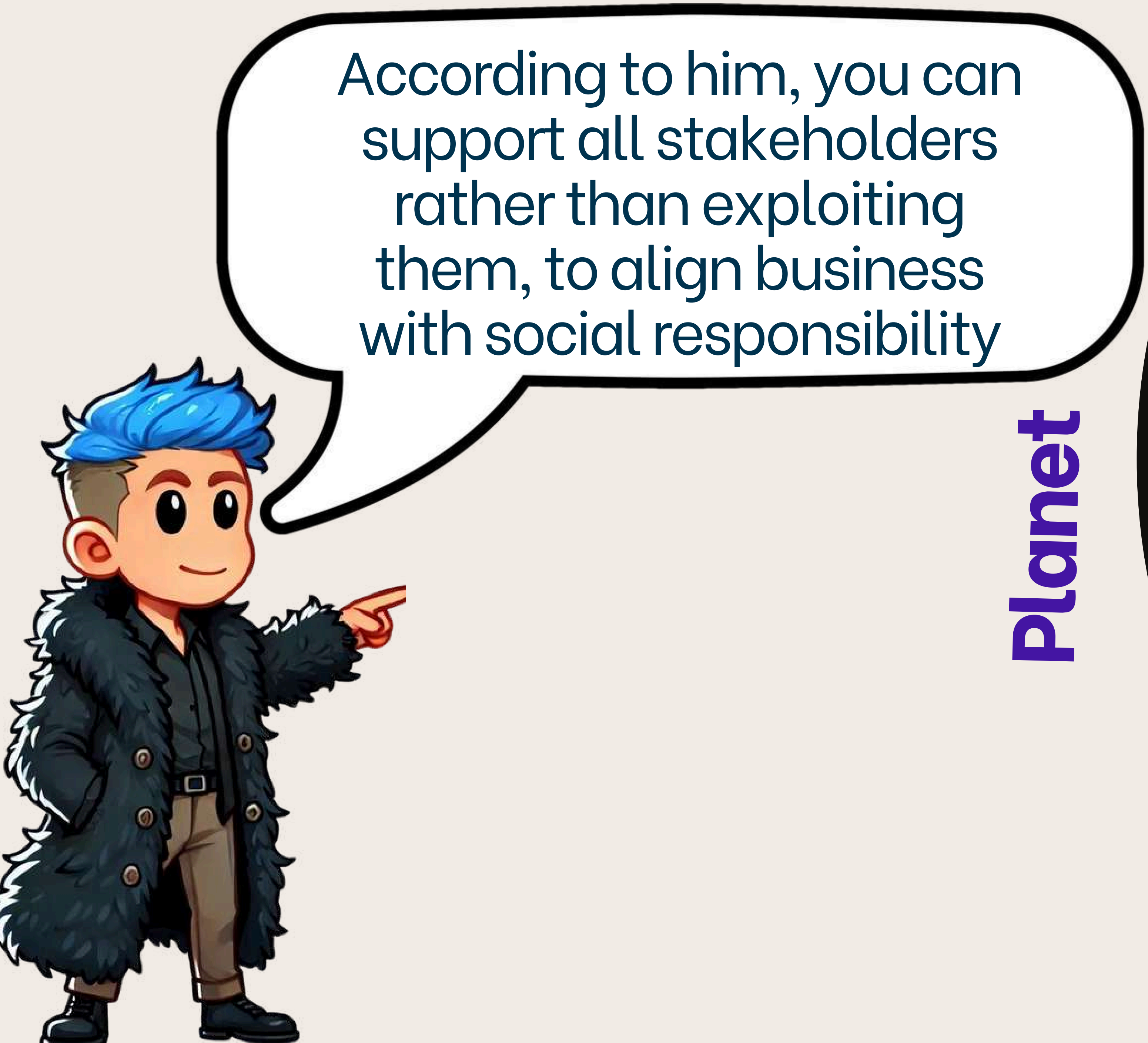
- In translation technology, extractivism extends beyond devices and networks, exploiting both the material resources and the unpaid labor of translators, who are often invisible behind high-tech solution

- In the context of climate change, translation technology should be as an integral part of the human ecosystem, with humans and technology co-acting and influencing each other.

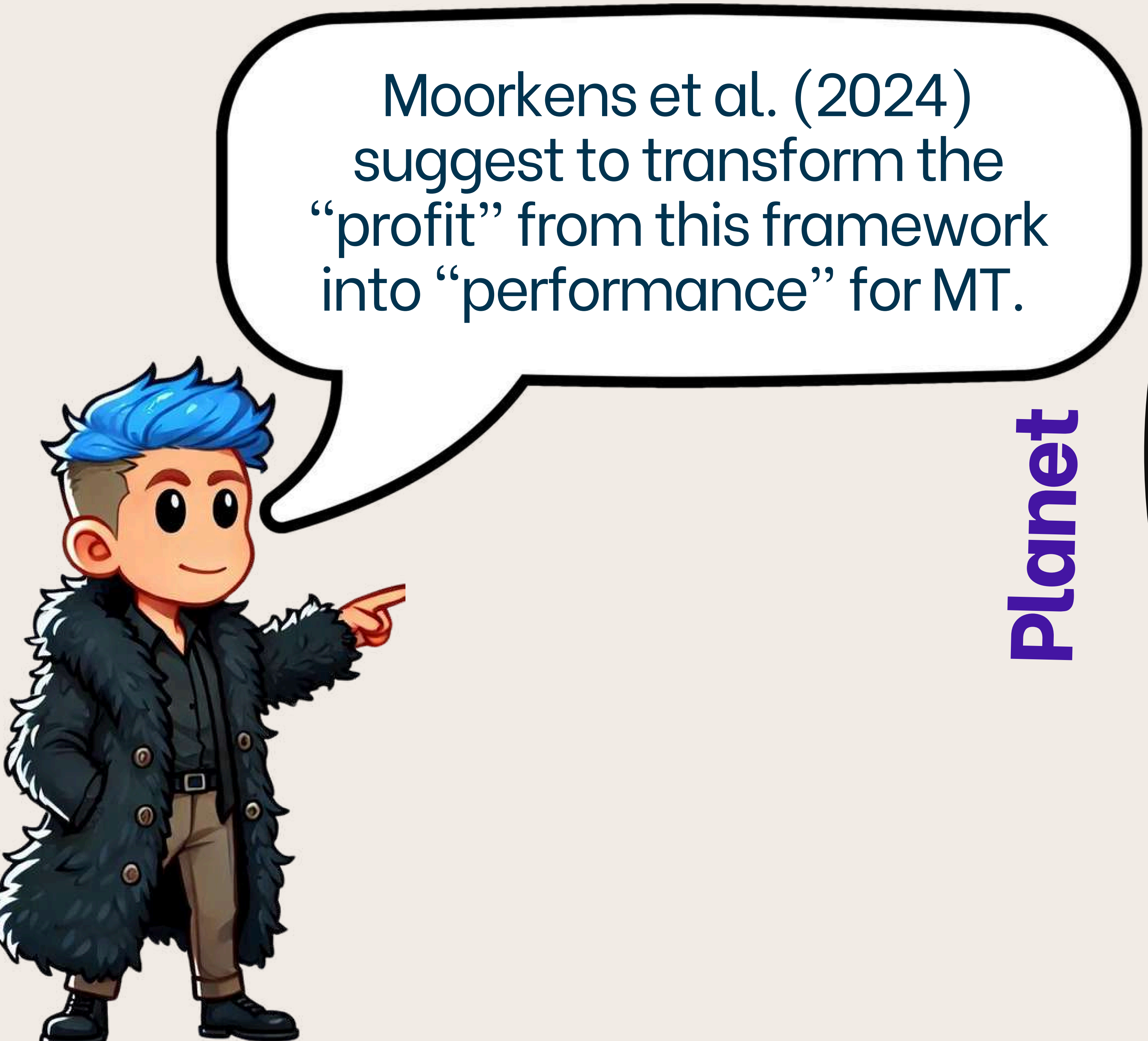
Towards a Triple Bottom Line (TBL)

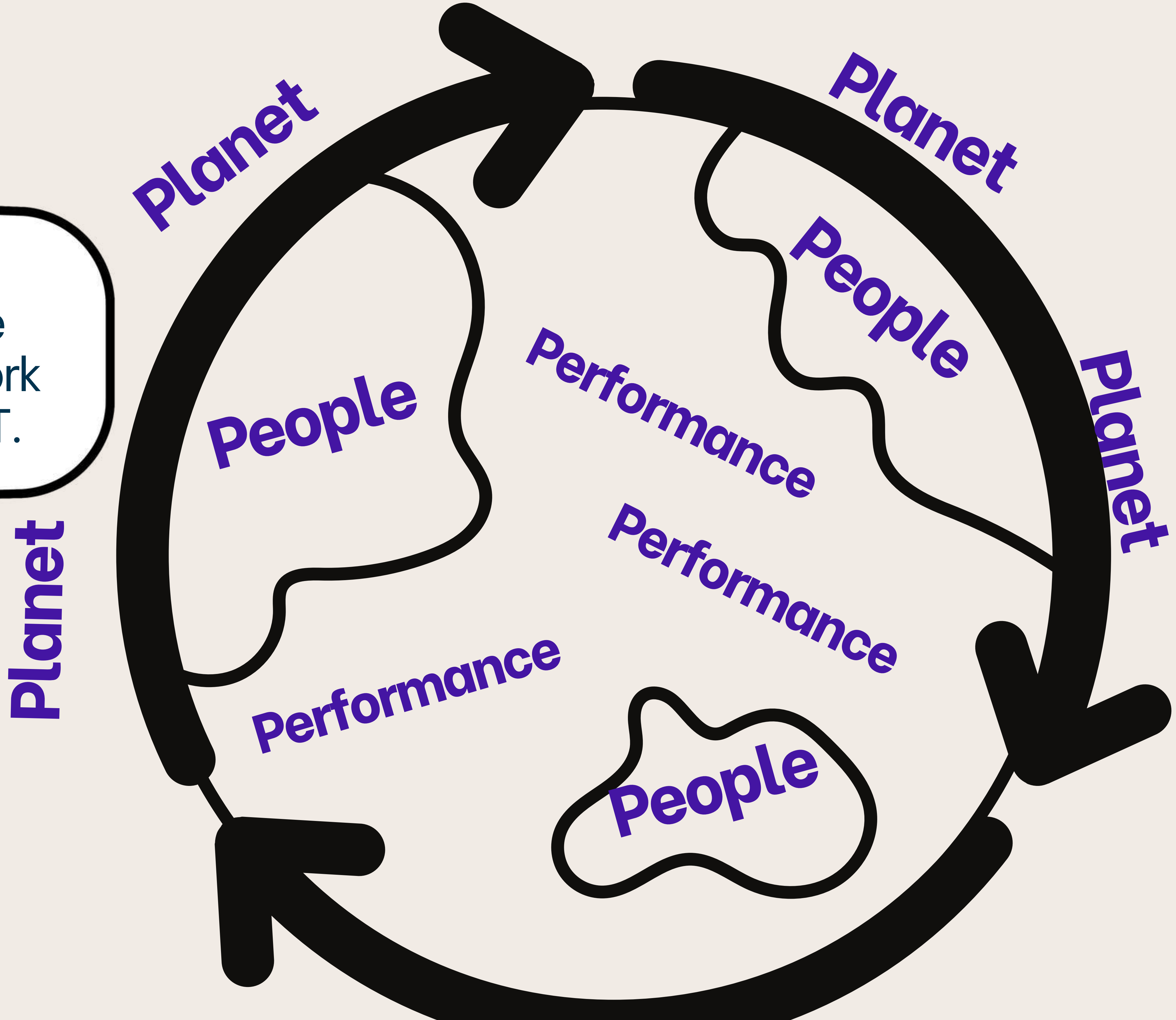
Elkington (1997) proposed a framework that included not just profit as a factor, but also people and the planet

Towards a Triple Bottom Line (TBL)



Towards a Triple Bottom Line (TBL)





Towards a Triple Bottom Line (TBL) - People

People

How to use automation to add value to People?

"Companies that focus on business sustainability are rewarded by reduced staff turnover and increased employee effort" (Wheelen et al. 2018)

Towards a Triple Bottom Line (TBL) - People

People

When we think of quality, let us think of quality for people.

"Reduced translation quality will introduce risk to users" (Moorkens et al. 2024, p. 8)

"When using automatic evaluation metrics that do not correlate well with human judgment, the use of those metrics should be called into question" (Moorkens et al. 2024, p. 8)

Towards a Triple Bottom Line (TBL) – People

People

When I acquire, train or pay for human data, am I being fair?

"Translators may have contractually agreed (or not) to allow their work to be repurposed for MT system training" (Moorkens et al. 2024, p. 8)

"The use of webcrawling for data acquisition is currently standard, without any real legal basis" (Moorkens et al. 2024, p. 8)

Towards a Triple Bottom Line (TBL) - Planet

Towards a Triple Bottom Line (TBL) - Planet

Planet

Automation technologies require energy, often referred to as compute costs

(Moorkens et al. 2024, p. 9)

Gen AI produces additional emissions with task-specific systems during training (Luccioni et al., 2023).

Training large machine learning models can emit as much CO2 as 1.5 cars over 20 years (Strubell et al., 2019).

Towards a Triple Bottom Line (TBL) - Planet

Planet

Machine learning can improve sustainability by optimizing resource use, reducing emissions, and increasing efficiency across various sectors. (Rolnick et al. 2023)

Technology leads to waste, and both manufacturing and disposal can pose risks

(Moorkens et al. 2024, p. 10)

We should move on from 'do no harm' to actively do good. (Moorkens et al. 2024, 10)

Towards a Triple Bottom Line (TBL) - Planet

Towards a Triple Bottom Line (TBL) - Planet

Let's look at performance now and reflect!

Performance

Use different types of metrics and report results accurately.

(Moorkens et al. 2024, p. 11)

Quality depends always on both context and situation (Drugan et al. 2018, p. 42)

We must be careful as to not exaggerate the capability of systems with sensasionalist terms (Moorkens et al. 2024, p. 11)

Towards a Triple Bottom Line (TBL) - Planet

There is a place for automatic metrics, where human evaluation is too slow or expensive. (Moorkens et al., 2024, p. 11)

Performance

Find the right context for automatic metrics

Investigate the weaknesses of automatic metrics, such as Armhein and Sennrich (2022) and Perrella et al. (2024).

(Moorkens et al. 2024, p. 11)

Towards a Triple Bottom Line (TBL) - Planet

Performance

Quality assessment matters for the safety of users

(Moorkens et al. 2024, p. 11)

Reduced quality introduces risk to the user and makes them put more effort in comprehension (Pym 2012)

Too much emphasis on performance and cost without attention to sustainability is not likely to bring long-term benefits (Moorkens et al. 2024, p. 11)

In this lecture you were able to...

Understand the strengths and limitations of automation and how it affects different types of end-users

Understand the strengths and limitations of automation on quality assessment and how you can make evaluations more robust and comprehensive

Understand how quality assessment can be performed with sustainability as an overarching factor encompassing people, planet and performance.

Thank you!
Questions?

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