Z-inspection®
Starting the Assessment process

Magnus Westerlund
Z-Inspection® Initiative
http://z-inspection.org

Arcada University of Applied Sciences, Helsinki Finland

Credits: **Roberto Zicari**

Structure of the Lesson

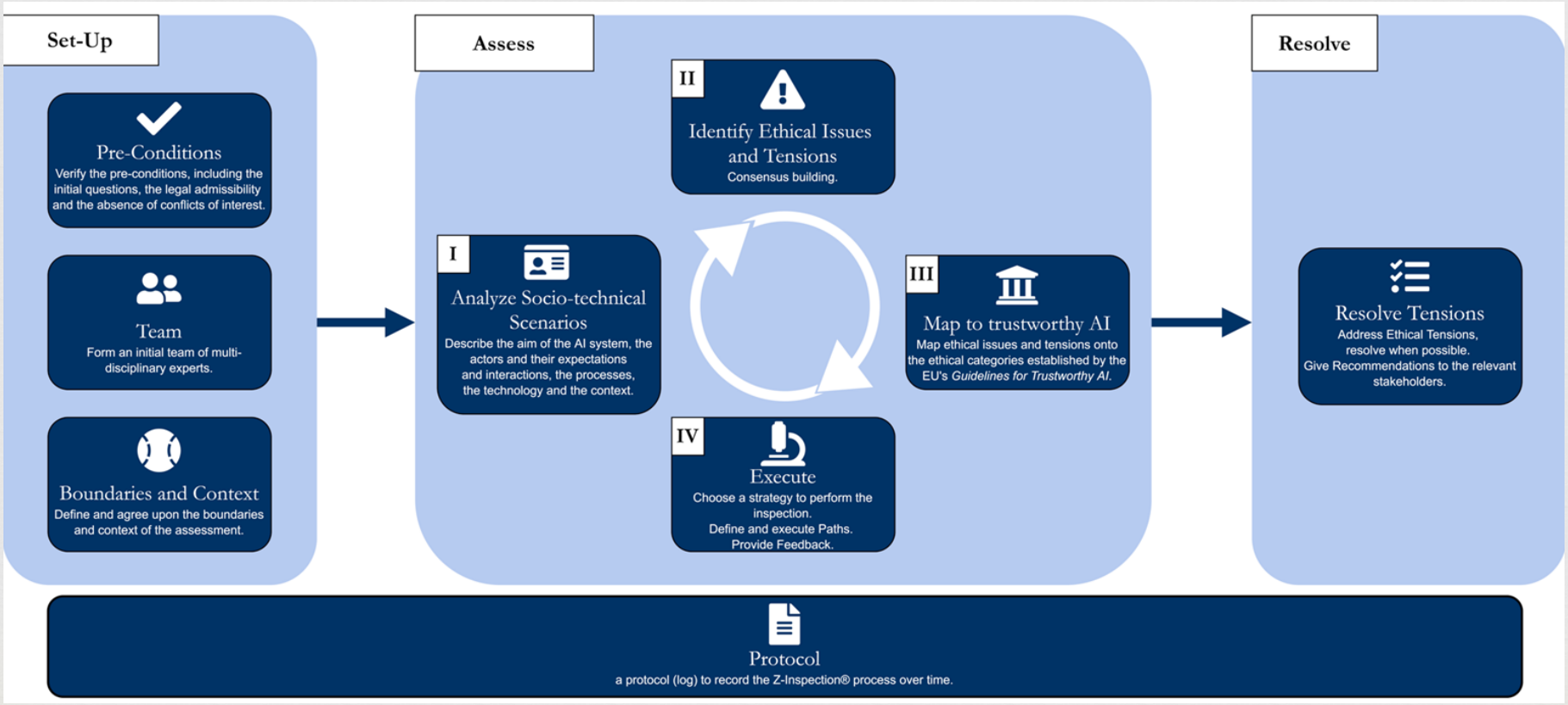
1. *Midterms*
2. *Background*
3. *Exploration of Claims*
4. *Path towards finding evidence*
5. *Simulated use case*

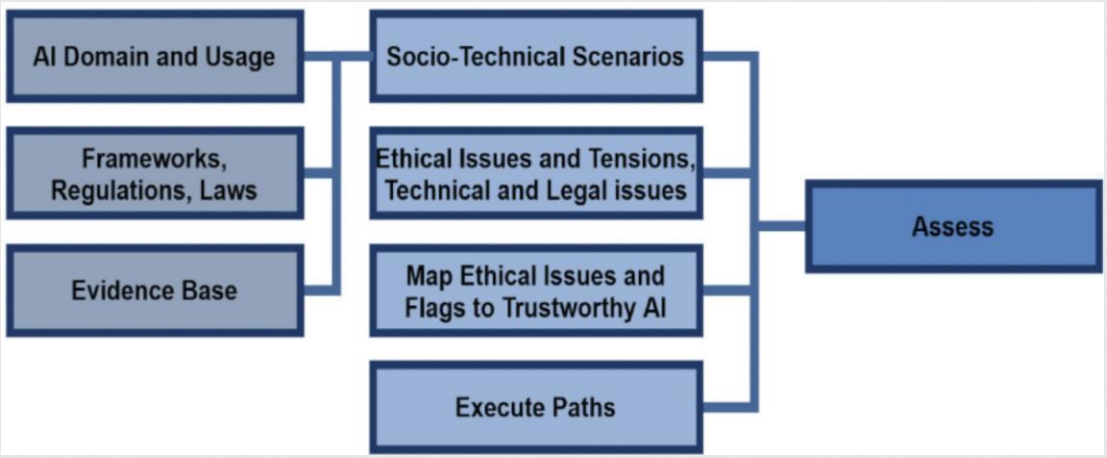
Midterms

- The goal of the mid-term report is to select an AI system (i.e., an AI product and or an AI-based service) used in healthcare, and start with the assessment process.
- Midterms will help you break down the problem.
- In teams of three to five students.

- Assess for “trustworthiness” based on the EU Ethics Guidelines for Trustworthy AI, adapted to the healthcare domain and the Z-inspection process.
- The Mid Term report should cover the following:
  - Define and agree upon the boundaries and context of the assessment.
  - Analyze Socio-technical scenarios
  - Identify Ethical Issues and Tensions
- Report must be delivered as slides and presented

Z-inspection® Process in a Nutshell





Socio-technical scenario building

Conceptual analysis aimed at exploring the AI System and its intended use

Scenario Building

- Contextualize the AI System to some real application
- Explore different perspectives to understand the (intended) use of the AI system.
- Develop an understanding of organizational processes and end-user interaction with the system

Claims, Arguments and Evidence

- **Concept building**: Mapping and clarifying ambiguities. Bridging disciplines, sectors, publics and cultures. Building consensus and managing disagreements.
- Understand technological capabilities and limitations. Build a robust **evidence base** to support **claims** and identify **tensions** (domain-specific). Understand the perspective of different members of society and different sub-disciplines.

We use Socio-technical Scenarios to learn about the intended use

By collecting relevant resources, a team of interdisciplinary experts create socio-technical scenarios and analyze them conceptually to describe:

- **the aim of the AI systems**,
- **the actors and their expectations and interactions**,
- **the process where the AI systems are used**,
- **the technology and the context (*ecosystem*)**.

Resulting in a number **of *issues* to be assessed**.

Socio-Technical Concept building

„An important obstacle to progress on the ethical and societal issues raised by AI-based **systems is the ambiguity of many central concepts currently used to identify salient issues**.„

- Terminological overlaps
- Differences between disciplines
- Differences across cultures and publics
- Conceptual complexity

- Source: *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Ethical Issues

Implemented technologies are never neutral, they always bring with them issues.

Develop an understanding of claims and evidence, to understand what ethical issues might persist.

Design considerations vs. ethical concerns

- An example of how to break down the AI system, above design considerations, and below concerns.

Raw Data	Features	Models	Inference
• Privacy • Completeness • Provenance	• Stereotyping • Importance	• Robustness • Quality • Fairness	• Explanation • Latency • Monitoring

Redress	Maintainability	Organisation	Systemic
• Appeal • Versioning • Trust	• Beneficence • Security • Backup	• Maturity • Accountability	• De-skilling • Multi-jurisdictional • Human Rights

Note, that this is a non-exhaustive list!

We develop an evidence base

This is an iterative process among experts with different skills and background with goal to:

- Understand technological capabilities and limitations
- Build a stronger evidence base to support claims and identify tensions (*domain specific*)
- Understand the perspective of different members of society

Source: Whittlestone, J et al (2019)

Claims, Arguments and Evidence (CAE)

Claims – "assertions put forward for general acceptance. They are typically statements about a property of the system or some subsystem.

Claims that are asserted as true without justification become **assumptions** and claims supporting an argument are called sub claims. "

- Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

Claims, Arguments and Evidence (CAE)

Evidence "that is used as the basis of the justification of the claim.

Sources of evidence may include the design, the development process, prior field experience, testing, source code analysis or formal analysis", peer-reviewed journals articles, peer-reviewed clinical trials, etc.

Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

Claims, Arguments and Evidence (CAE)

Arguments link the evidence to the claim.

They are defined as Toulmin's warrants and are the "statements indicating the general ways of arguing being applied in a particular case and implicitly relied on and whose trustworthiness is well established", together with the validation for the scientific and engineering laws used.

Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

Develop an evidence base

- Technology is generally designed for a highly specific purpose, however, **it is not always clear what the technologies unintended harm might be**.

- Therefore, an important part of our assessment process is **to build an evidence** base through the **socio-technical scenarios to identify tensions as potential ethical issues.**

Claims

- "AI developers regularly make **claims regarding the properties of AI systems they develop as well as their associated societal consequences.** "

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims https://arxiv.org/pdf/2004.07213.pdf

Identify Claims

- **Claims** for technological capability (for example aim, performance, architecture, or functionality, etc.) serve as an important input in developing the **evidence base.**
- This is an iterative process among experts of the assessment team with different skills and backgrounds with a goal to understand technological capabilities and limitations

Verifiable Claims

- „**Verifiable claims** are statements for which **evidence** and **arguments** can be brought to bear on the likelihood of those claims being true.

- The degree of attainable certainty in such claims will vary across contexts. „

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims https://arxiv.org/pdf/2004.07213.pdf

Examples of Claims

- We will *adhere* to the data usage protocols we have specified;
- The cloud services on which our AI systems run are *secure*;
- We will *evaluate* risks and benefits of publishing AI systems in partnership with appropriately qualified third parties;

Examples of Claims

- The AI system *is very accurate*…
- The AI system is *more accurate then*….
- The AI system is *98% accurate*...
- The AI *predicts with high quality* ….
- *Using the AI system results in saving XXX do*llars…

Examples of Claims

- We *will not create or sell AI systems that are intended to cause harm;*
- We *will assess and report any harmful societal impacts of AI systems that we build;* and
- Broadly, *we will act in a way that aligns with society's interests.*

“Keep your AI claims in check” USA Federal Trade Commission Division

Are you exaggerating what your AI product can do?

Or even claiming it can do something beyond the current capability of any AI or automated technology?

For example, we’re not yet living in the realm of science fiction, where computers can generally make trustworthy predictions of human behavior.

Your performance claims would be deceptive if they lack scientific support or if they apply only to certain types of users or under certain conditions.

- Source: Keep your AI claims in check, By Michael Atleson, Attorney, US Federal Trade Commision Division of Advertising Practices February 27, 2023

“Keep your AI claims in check” USA Federal Trade Commission Division

Are you promising that your AI product does something better than a non-AI product?

It’s not uncommon for advertisers to say that some new-fangled technology makes their product better – perhaps to justify a higher price or influence labor decisions. You need adequate proof for that kind of comparative claim, too, and if such proof is impossible to get, then don’t make the claim.

Source: Keep your AI claims in check, By Michael Atleson, Attorney, US Federal Trade Commision Division of Advertising Practices February 27, 2023

“Keep your AI claims in check” USA Federal Trade Commission Division

Are you aware of the risks?

You need to know about the reasonably foreseeable risks and impact of your AI product before putting it on the market. If something goes wrong – maybe it fails or yields biased results – you can’t just blame a third-party developer of the technology. And you can’t say you’re not responsible because that technology is a “black box” you can’t understand or didn’t know how to test.

- Source: Keep your AI claims in check, By Michael Atleson, Attorney, US Federal Trade Commision Division of Advertising Practices February 27, 2023

“Keep your AI claims in check” USA Federal Trade Commission Division

Does the product actually use AI at all?

If you think you can get away with baseless claims that your product is AI-enabled, think again. In an investigation, FTC technologists and others can look under the hood and analyze other materials to see if what’s inside matches up with your claims. Before labeling your product as AI-powered, note also that merely using an AI tool in the development process is not the same as a product having AI in it.

- Source: Keep your AI claims in check, By Michael Atleson, Attorney, US Federal Trade Commision Division of Advertising Practices February 27, 2023

Building a solid knowledge / evidence base

- We suggest **building a solid knowledge / evidence base** among all team members of the use case before the inspection starts and also a solid Q&A log during the inspection process.
- Experts may approach the use case quite differently:
- Interpretations of and expectations for the AI tool being inspected may differ
- Focus of interest may be very different

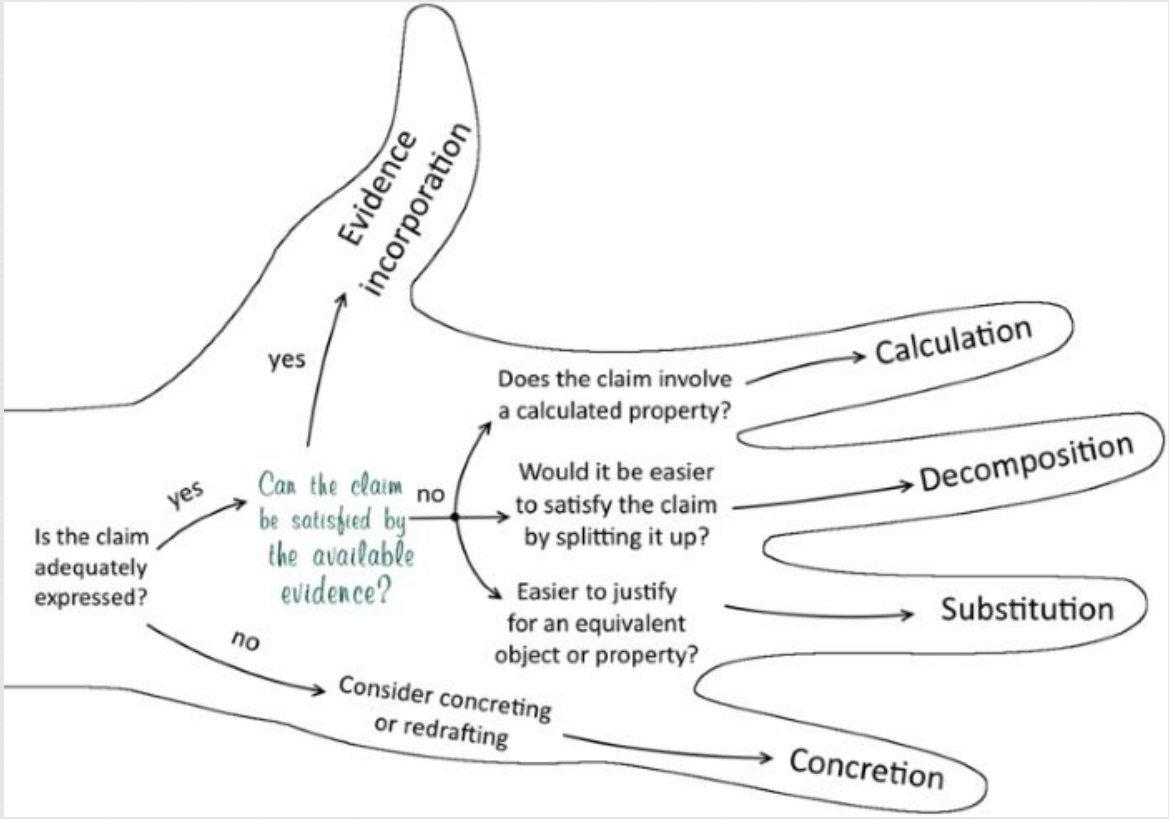
Claims, Arguments and Evidence

- **The claims, arguments and evidence (CAE) framework (*)** can help with the structuring of the use case in a clear and precise form that is supported by evidence.
- For example, each of the claims should be about only one specific property of the system and at the same time, it should be phrased in a way that is clearly verifiable or falsifiable. The CAE framework also provides guidance on how to disseminate complex claims into easier ones .

(*) https://claimsargumentsevidence.org

Claims, Arguments and Evidence

Source: https://claimsargumentsevidence.org

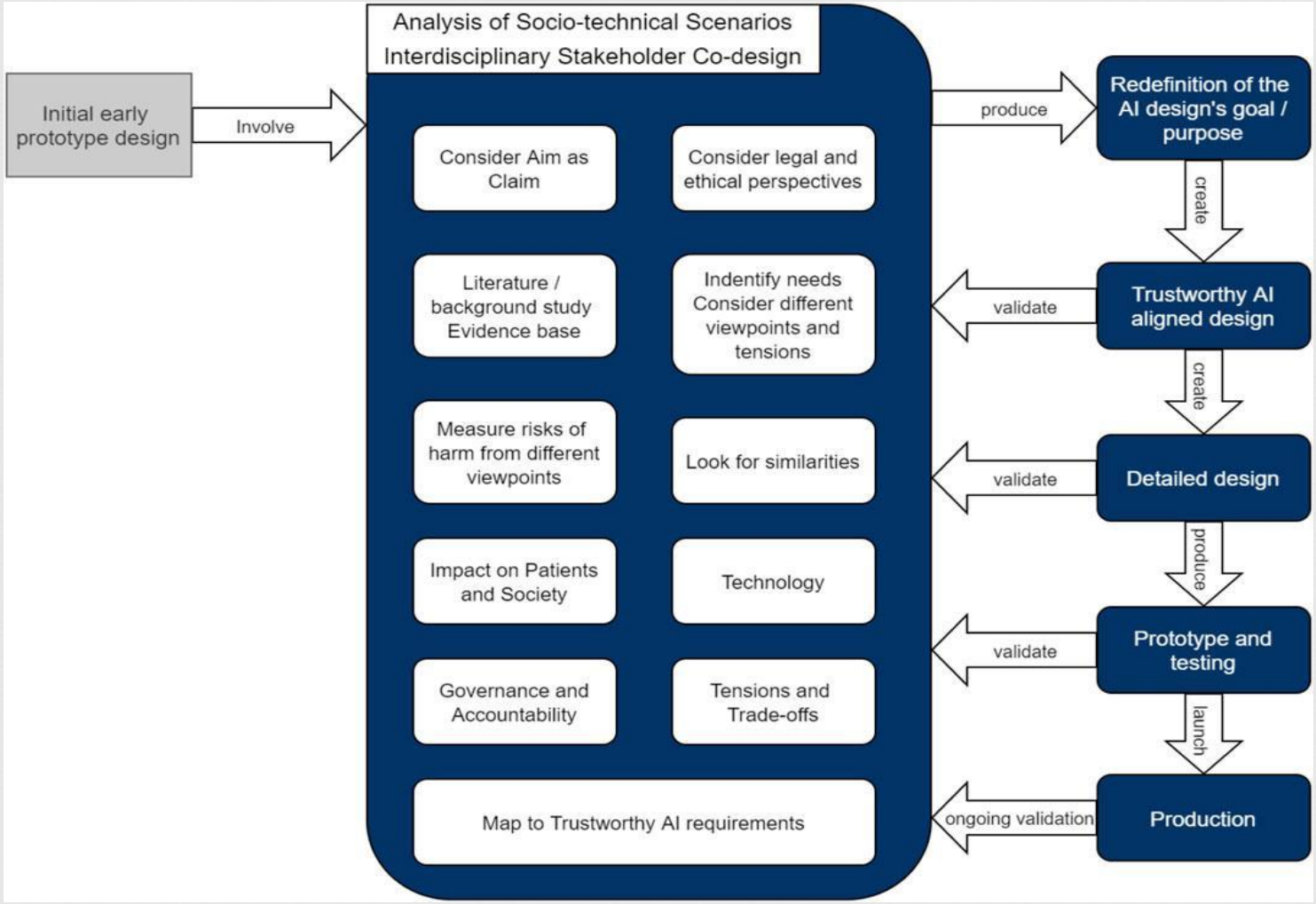


The Z-Inspection® process: Co-Design

In **design and development phases:**

Z- Inspection® can be used as a **co-creation process** to help AI engineers, domain experts to ensure that the design of their AI system meets the trustworthy AI criteria.

Co-Design process



When in Co-design.

- **Consider the AI *initial design as a Claim that needs to be verified with evidence.***

- Example: When designing, training and testing an AI-system (e.g. Machine-Learning algorithm) we do "embed" into the system notions such as "**good**", "**bad**", "**healthy**", "**disease**", etc. mostly not in an explicit/transparent way.

Hypothetical case

Project Backstory: DialogueEducator - Nurturing Metacognitive Learning in Higher Education

- In the heart of the Basque Country, a revolution in higher education pedagogy is brewing. Here, we understand that the traditional methods of teaching no longer suffice in preparing students for the dynamic challenges they will face in the ever-evolving professional world. The key? Metacognition - or, simply put, learning how to learn.
- Introducing DialogueEducator:
- DialogueEducator is our pioneering project, designed to catapult higher education into a new era of learning. Utilizing the advanced AI capabilities of ChatGPT, we have developed a sophisticated dialogue tool tailored specifically to the unique linguistic and cultural context of the Basque Country. But DialogueEducator is far more than a mere dialogue tool; it is an interactive workbook, a personal tutor, and most importantly, a mirror through which students can understand their own learning processes.
- Our journey begins with a course on Natural Language Processing (NLP) for Computer Science students. In this course, DialogueEducator will serve as a dynamic, interactive platform that not only imparts knowledge but also engages students in a reflective learning process, helping them understand their own cognitive mechanisms and learning preferences.
- Functional Requirements Derived from Metacognitive Aims:
- Interactive Learning Modules: DialogueEducator must feature a series of interactive learning modules that use real-time dialogue to present course material. These modules should be designed to adapt to students' responses, encouraging active participation and reflection on the learning process.
- Self-assessment Tools: The platform should include integrated self-assessment tools that prompt students to evaluate their understanding of the course material. This feature should provide immediate feedback, allowing students to identify areas of strength and weakness and to track their progress over time.
- Resource Recommendation System: Based on self-assessment results and interactive dialogue, DialogueEducator should suggest additional resources tailored to students' individual learning needs. This system must be dynamic, adjusting recommendations based on students' evolving understanding and preferences.
- Reflection Prompts: Throughout the learning modules, the tool should present students with targeted reflection prompts designed to encourage metacognitive thinking. These prompts should guide students to consider their learning strategies and outcomes critically, fostering a habit of continuous self-reflection and improvement.
- Strategy Planning and Monitoring: DialogueEducator should facilitate the development of personalized learning strategies. This involves setting specific, measurable, achievable, relevant, and time-bound (SMART) goals, planning learning activities, and regularly reviewing progress.
- Multi-Lingual Support: Given the cultural context of the Basque Country, the tool must support bilingual or multilingual options, ensuring accessibility and inclusivity for all students.
- Data Security and Privacy: To safeguard student data, the platform must adhere to stringent data security and privacy regulations. This involves secure data storage, encrypted data transmission, and strict adherence to GDPR and other relevant legal frameworks.

Design considerations vs. ethical concerns

- An example of how to break down the AI system, above design considerations, and below concerns.

Raw Data	Features	Models	Inference
• Privacy • Completeness • Provenance	• Stereotyping • Importance	• Robustness • Quality • Fairness	• Explanation • Latency • Monitoring

Redress	Maintainability	Organisation	Systemic
• Appeal • Versioning • Trust	• Beneficence • Security • Backup	• Maturity • Accountability	• De-skilling • Multi-jurisdictional • Human Rights

Note, that this is a non-exhaustive list!

Claims by a hypothetical tool Owner

- Addressing the ethical concern of "**Data Completeness and Quality**" requires a meticulous approach, especially when considering the development of an AI-driven educational tool like DialogueEducator. Below are detailed claims that outline how we plan to address these concerns at the raw data design step:

- Comprehensive Data Collection Protocols:
- Claim: Our data collection protocols are standardized and methodical, ensuring uniformity across various sources. This uniformity mitigates the risk of systemic biases that might occur due to discrepancies in data collection methods.
- Technical Details: We employ advanced ETL (Extract, Transform, Load) processes that are configured to identify, flag, and manage anomalies at the data extraction stage. These processes are automated to ensure consistency and are subject to regular audits.

- Robust Data Preprocessing Techniques:
- Claim: We utilize state-of-the-art data preprocessing techniques to cleanse the data, handling missing values, outliers, and noise that could potentially skew the model's performance and decision-making fairness.
- Technical Details: Techniques include statistical methods for outlier detection (like Z-score, IQR methods), data imputation methods for handling missing data (e.g., mean/mode/median imputation, predictive models, or using algorithms like KNN, which inherently handle missing data), and noise reduction techniques (like data smoothing, filtering, or ensemble methods).

- Systematic Bias Identification and Mitigation:
- Claim: We are committed to identifying and mitigating any form of systematic bias during data collection and preprocessing to ensure fair representation and treatment of all user groups.
- Technical Details: This involves using fairness metrics and tools (like Fairlearn or IBM's AI Fairness 360) to assess and mitigate potential biases in the data. Additionally, we perform stratified sampling to ensure diverse and representative data subsets, which help in understanding and addressing inherent biases.

- Continuous Data Quality Monitoring:
- Claim: Our system doesn't just assume a one-time data cleaning process is sufficient. Instead, we have established continuous data quality monitoring to promptly identify and address data quality issues that might emerge over time.
- Technical Details: Continuous monitoring involves automated data quality checks using predefined quality rules (like data validation frameworks), real-time anomaly detection systems, and routine data audits. These processes are facilitated by machine learning operations (MLOps) tools that provide continuous tracking and automated workflows for data validation.

- Transparent Data Documentation and Provenance:
- Claim: We maintain comprehensive documentation of our data sources, collection methodologies, and preprocessing decisions to ensure transparency and traceability. This practice is crucial for external reviews and for users who have concerns about the origin and handling of the data influencing the DialogueEducator.
- Technical Details: Data dictionaries, detailed logs of data transformation and cleaning steps, and metadata management practices are maintained. We employ data versioning tools to track changes and ensure that each dataset's history is fully documented and auditable.

- User Consent and Ethical Data Collection:
- Claim: User consent is paramount in our data collection process. We ensure that all our data is ethically sourced, with clear communication to users about what data is being collected and why it is necessary for the functioning of DialogueEducator.
- Technical Details: This involves GDPR-compliant consent management processes, clear and accessible privacy policies, and user-friendly interfaces for consent management. We use secure sessions and encryption to protect user data during and after collection.

- Inclusive Data Representation:
- Claim: Recognizing the diverse user base of DialogueEducator, we commit to data collection strategies that ensure inclusive representation of various demographic groups, thereby enhancing the tool's fairness and effectiveness.
- Technical Details: This involves conducting demographic analysis of our user base, consulting with diversity and inclusion experts, and potentially using oversampling or undersampling strategies to address imbalances in representation.

Hypothetical case

Potential Ethical Issues

1. Comprehensive Data Collection Protocols:

The use of advanced ETL processes to identify and manage anomalies is commendable. However, to ensure the preservation of Basque heritage, it's crucial to ensure that the sources of data are culturally relevant and appropriate for the Basque context. The consistency across various sources is essential, but the nature of these sources is equally pivotal.

2. Robust Data Preprocessing Techniques:

The techniques mentioned for data preprocessing are technically sound. Given my statistical background, I appreciate the attention to detail in handling outliers, missing data, and noise. However, I'd like to see more information on how these techniques are tailored for the Basque linguistic and cultural nuances.

3. Systematic Bias Identification and Mitigation:

Using fairness metrics and tools is a positive approach, and stratified sampling ensures diverse representation. Nevertheless, it's essential to understand the criteria for stratification, especially in the context of the Basque Country, to ensure that no subgroups are marginalized.

4. Continuous Data Quality Monitoring:

Continuous monitoring is an excellent practice. Given the dynamic nature of language and culture, this will be especially valuable for the Basque context. However, I'd emphasize the importance of periodic human reviews to ensure cultural sensitivity.

5. Transparent Data Documentation and Provenance:

Transparency is crucial. Maintaining comprehensive documentation will be essential for trust. I would recommend that the team also incorporates feedback mechanisms where users and experts can flag and discuss potential issues related to data and its cultural significance.

6. User Consent and Ethical Data Collection:

Ethical data collection is paramount. While GDPR compliance is essential, it's also necessary to ensure that users from the Basque region understand the implications of their data being used, given the historical context of the Basque people's suppression.

7. Inclusive Data Representation:

Consulting with diversity and inclusion experts is a step in the right direction. But considering the unique cultural and linguistic context of the Basque Country, I recommend collaborating with local cultural organizations and linguists to ensure genuine inclusivity.