onetrust

Let’s do it live: Role-playing a GenAI project risk assessment

19 January 2024

Speakers

Henri Kujala

Global Head of Privacy by Design

Vodafone

Linda Thielova

Head of Privacy Center of Excellence and DPO

OneTrust

Bex Evans

Sr. Product Marketing Manager, Responsible AI

OneTrust

Risks of using Large Language Models

Bias, Discrimination, Hate speech

Misinformation

Information Security & Privacy

Malicious & Fraudulent use

Environmental & socioeconomic harms

Scenario:

A retailer of beauty, wellness, and health products looks to establish a new sales channel using a GenAI powered chatbot.

A risk-based approach starts with understanding inputs and outputs

AI Governance

Inputs → Model → Outputs

Data sets

- Product descriptions, including medication information
- Customer inquiries
- Health information
- Billing information
- Contact information

Black box

Outcomes

- Product recommendations
- Customer profiling
- Order fulfillment

How do you start assessing risk?

Business risk

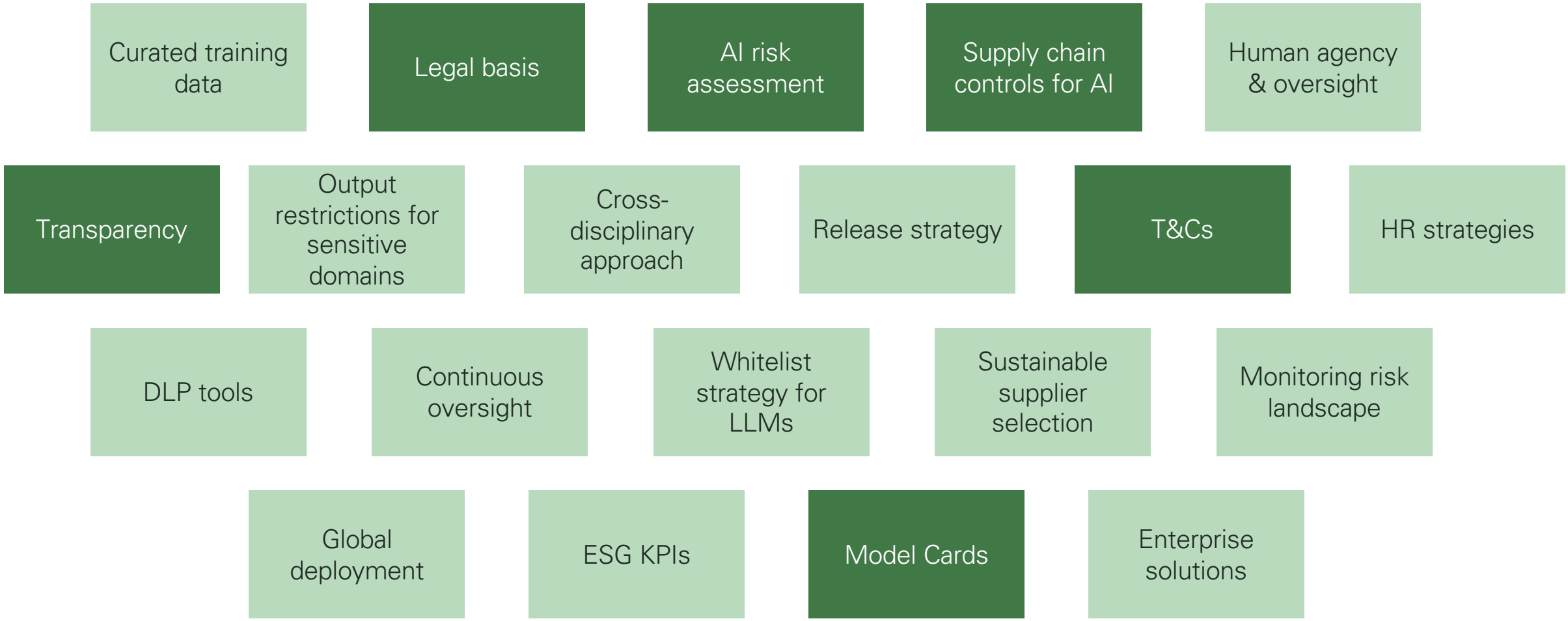
Technology risk

Ethical risk

Legal risk

What practical actions items can be applied to mitigate the risks?

Mitigating controls for different risks



What are model cards?

A concise document that provides key information about a machine learning model, promoting transparency and responsible AI usage.

Model Details:

- Name, version, and model type (e.g., neural network, decision tree).
- Intended use cases.

Model Architecture:

- Description of the model's structure, including layers and activation functions.

Training Data and Methodology:

- Dataset details, such as size, sources, and preprocessing.
- Information about training techniques, optimizers, loss functions, and hyperparameters.

Performance Metrics:

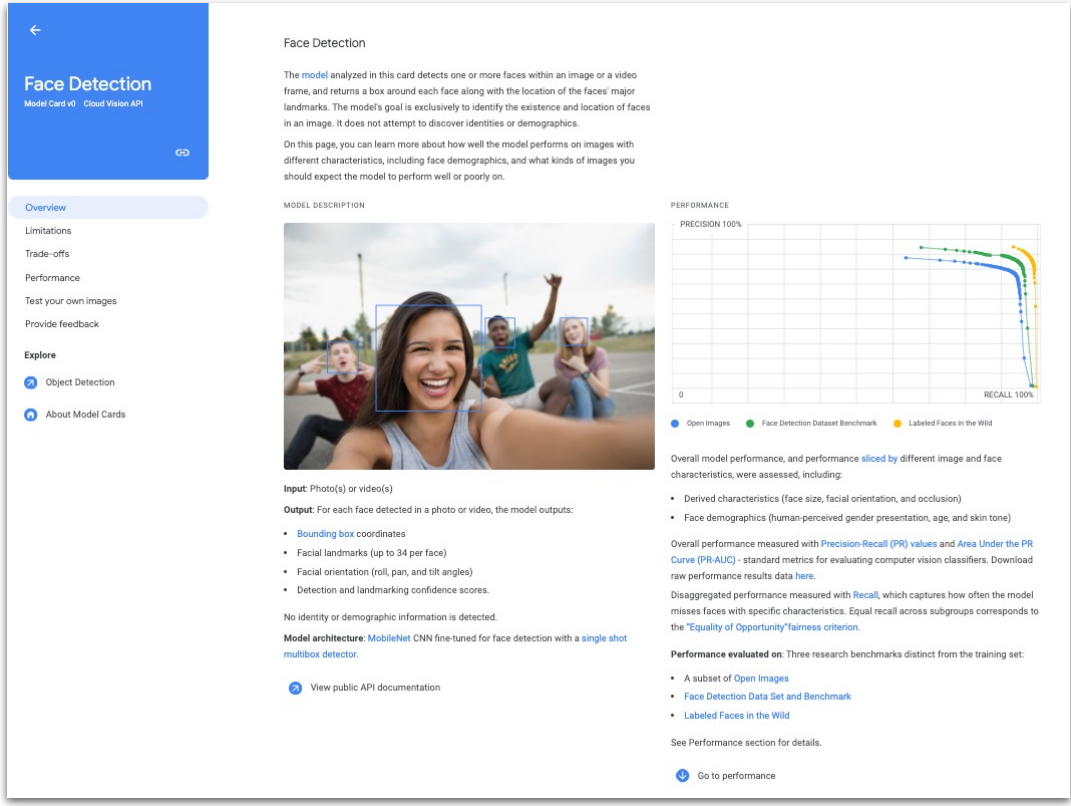
- Model's performance on various metrics, e.g., accuracy, precision, recall, F1 score.
- Performance across different subsets of data.

Potential Biases and Limitations:

- Addressing potential biases and limitations, such as imbalanced data and overfitting.
- Generalization capabilities and suitability for specific use cases.

Responsible AI Considerations:

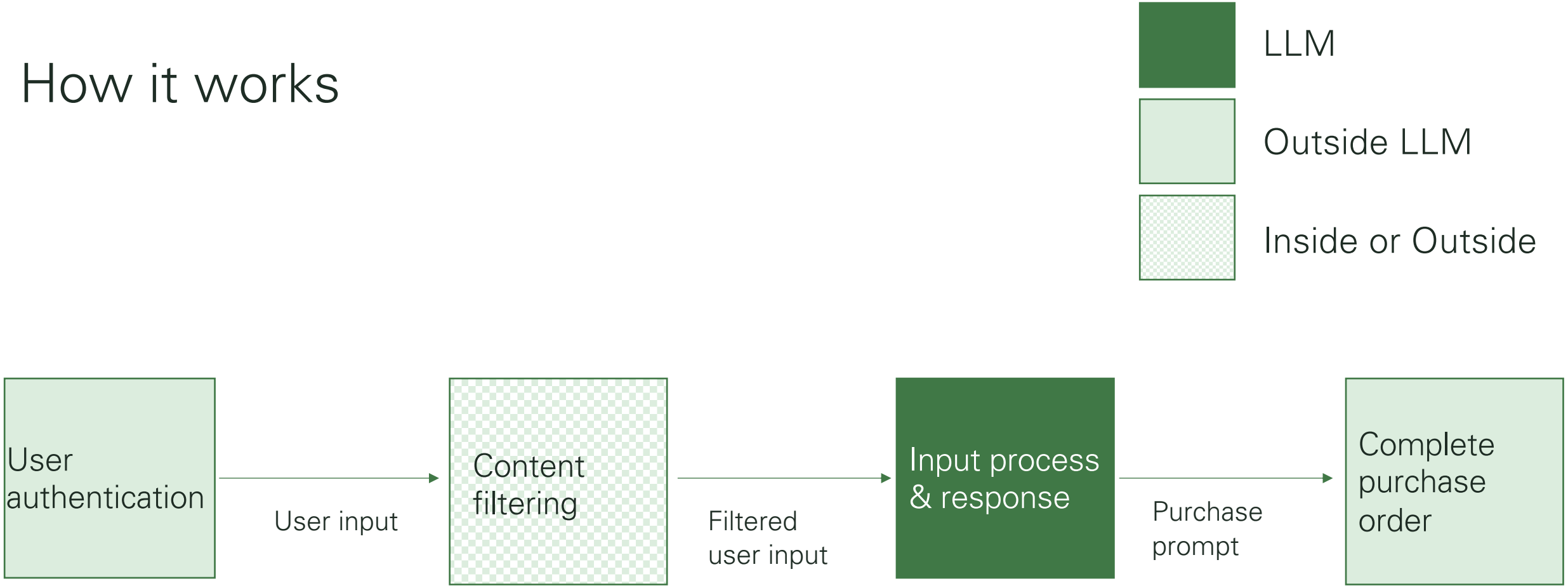
- Ethical and societal implications.
- Privacy, fairness, and transparency considerations.
- Recommendations for further testing, validation, or monitoring.



Google, Face Detection Model Card v0

Guidance on chatbots

How it works



Closing the loop

Model approval

Data minimization

Exposure of LLM features

Human oversight

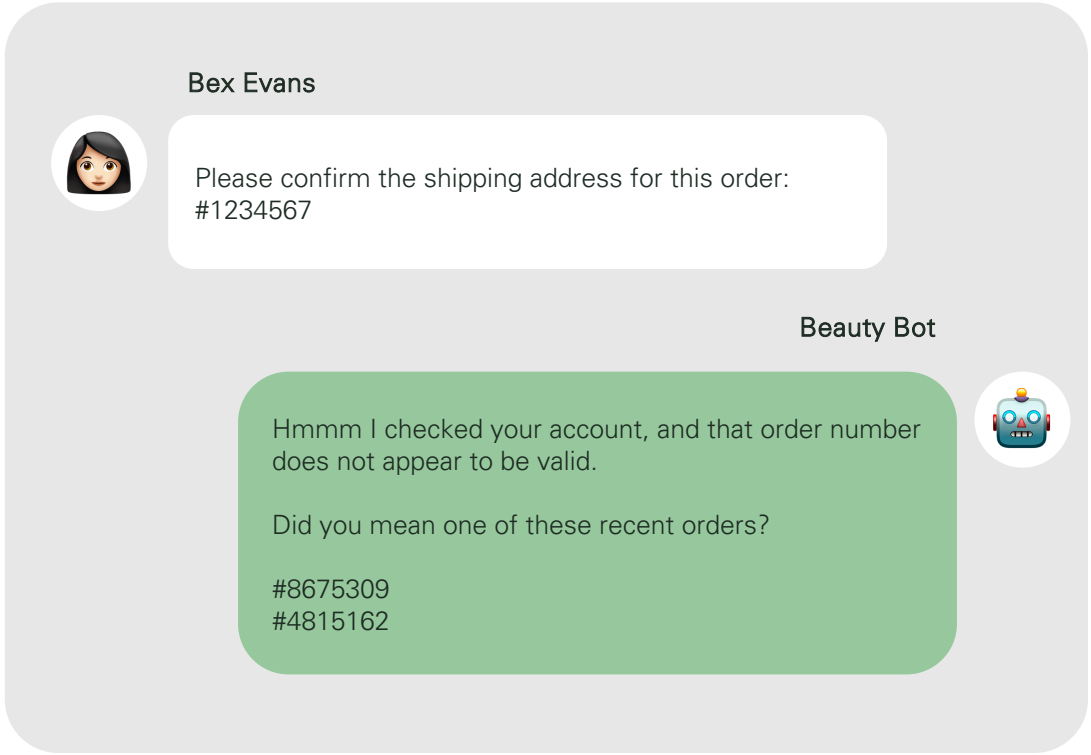
Acceptable use

Authorization / authentication

Test scenarios

Red teaming

Chaining LLMs



Questions

Thank You

Connect with us on LinkedIn

onetrust

Bex Evans, OneTrust

Henri Kujala, Vodafone

Linda Thielova, OneTrust