Designing an Ethical Framework

Rashika Dabas (101458141) (Applied A.I. Solutions Development Program)

Dr. Cindy Gordon

December 12, 2023

Abstract

This research proposes a framework for ethical AI development and application in the fashion industry, focusing on the use case of virtual try-on (VTO) technology. It draws inspiration from existing frameworks like the Ethical 8 Risk Zone Framework, IBM AI Explainability Toolkits, and Google AI Fairness Toolkit and emphasizes five key principles: fairness, transparency, explainability, robustness, and privacy. The framework aims to mitigate potential ethical risks associated with VTO, such as bias, manipulation, and data misuse, and ensure its responsible and beneficial use for both consumers and businesses.

Introduction

Firstly, let me talk about the most important step of the 7-step plan for operationalizing data and AI ethics. To **create a data and AI ethical risk framework** tailored to any specific industry, it should include, at a minimum, articulation of ethical standards (clear statement of the company's ethical values and principles, as well as a list of potential ethical nightmares that the company wants to avoid), identification of stakeholders (list all of the relevant stakeholders who could be affected by the company's use of data and AI, both internal and external), governance structure (outline the roles and responsibilities of different individuals and groups within the company with respect to data and AI ethics), quality assurance program (design a process for monitoring and evaluating the effectiveness of the company's data and AI ethics program) and integration with operations (Ensure that ethical considerations are built into all aspects of the company's data and AI development and deployment processes).

Frameworks

I have considered the following **3 frameworks** for **Rashika's Ethical AI Framework**:

1. **The Ethical 8 Risk Zone Framework** developed by the Institute for the Future (IFTF) states 8 key areas where emerging technologies, particularly AI, have the potential to cause harm and create ethical dilemmas.
2. **IBM AI Explainability Toolkits**, also known as AI Explainability 360, is an open-source software toolkit that helps developers and data scientists understand and explain the predictions made by machine learning models.
3. The **Google AI Fairness Toolkit**, a suite of open-source tools and resources designed to help developers and researchers build and deploy fair and unbiased machine learning models, offers various functionalities under three main categories: bias detection, fair model training, and model evaluation and interpretation.

Use Case

I've decided the use case to be **virtual try-on (VTO) in the fashion industry**. Virtual try-on technology is revolutionizing the fashion industry, allowing customers to virtually "try on" clothing using augmented reality (AR). This has numerous benefits for consumers, primarily convenience, personalization, reduced decision fatigue (which leads to more confident purchase decisions considering the photogenic look), and sustainability (which reduces the

need for physical returns and samples). Businesses experience increased sales, reduced return rates, and improved product development (by using data on how customers interact with clothing to improve product design and fit). Despite its benefits, VTO also raises several **AI ethics considerations**, such as:

- Bias and discrimination: VTO algorithms can be biased based on the data they are trained on. This can lead to inaccurate or discriminatory results, particularly for people of colour, different body types, or with disabilities.
- Privacy concerns: VTO technology often requires users to provide personal data, such as body measurements and photos. This data must be collected and stored securely and ethically.
- Transparency and explainability: It is important for consumers to understand how VTO algorithms work and how their data is being used. Algorithms should be transparent and explainable to ensure fairness and accountability.
- Accessibility: VTO technology should be accessible to everyone, regardless of their technical skills or physical abilities.

Principles

By addressing the above AI ethics concerns, VTO can become a powerful tool for the fashion industry that benefits every individual involved in the process. It is important to use this technology responsibly and ethically to ensure a fair and inclusive future for fashion.

This leads to the **5 key principles/criteria** of **Rashika's Ethical AI Framework** for the chosen use case.

1. Fairness:

   The Machine Ethics and Algorithmic Bias Zone of the **Ethical 8 Risk Zone Framework** focuses on the potential for AI algorithms to be biased, discriminatory, and unfair. Another zone, namely economic and asset inequalities, highlights the potential for AI to exacerbate existing economic inequalities and create a society where the rich get richer, and the poor get poorer.

   IBM Explainability Tools also mentions that AI models should not discriminate against any individual or group based on protected characteristics such as race, gender, religion, or disability and can help identify and address bias in AI models. For example, they can help identify which features are contributing the most to bias and how to mitigate their impact. This can be handled using diverse datasets that represent a wide range of people to train VTO algorithms and implementing regular testing for bias and discrimination.

   Moreover, the **Google AI Fairness Toolkit** supports bias detection through:

   - Metrics: Provides various metrics for quantifying different types of bias in datasets and models, including fairness metrics like equalized odds ratio and true positive rate difference, as well as data bias metrics like feature parity and label distribution divergence.
   - Visualization: Offers tools for visualizing data and model bias, such as counterfactual examples and fairness-aware data exploration tools.

and fair model training via:

- Effective Data Preprocessing: Provides techniques for mitigating bias in data, such as data balancing and feature selection algorithms.
- Fairness-aware Model Training: Offers algorithms and techniques for training machine learning models that are fairer, including algorithms like calibrated equalized odds and constrained fairness optimization.
- Fairness Monitoring: Offers tools for monitoring the fairness of deployed models over time and detecting any emerging bias issues.

Also, it facilitates compliance with regulations in organizations that require fairness in models.

With respect to **VTO**, there are some concerns about its fairness. These algorithms may be more likely to show certain types of clothes to certain types of people. For example, an algorithm might be more likely to show a white person wearing a white dress than a black person wearing a white dress. This type of bias can lead to consumers feeling excluded or discriminated against. One important step to nip this in the bud is to make sure that the data used to train these algorithms is representative of the entire population. Another important step is to curb biasedness, if there is any, against certain groups of people.

2. Transparency:

The Truth, Disinformation, and Propaganda Zone of the **Ethical 8 Risk Zone Framework** highlights the potential for AI to be used to spread misinformation and propaganda, manipulate public opinion, and undermine trust in institutions. A simple mitigation strategy for this is to provide clear privacy policies explaining how user data is collected, used, and stored.

Also, **IBM Explainability Tools** are open-source and community-driven (available on GitHub), allowing for better contributions from the community, which fosters transparency and innovation. They can help improve the transparency of AI models and build trust with users. For example, they can provide explanations that are specific to individual predictions and that are based on features that are relevant to the task. By providing information about the data used to train models, the algorithms used to make predictions, and the factors that influence model decisions, there's more transparency around the development and deployment, and individuals are accountable for the decisions made by AI systems. Through this, companies can build trust with users and stakeholders.

The **Google AI Fairness Toolkit** is also open-source and freely available, which makes it accessible and transparent to anyone wanting to build AI models. It fosters transparency and trust by enabling developers and users to understand how models make decisions and identify potential biases.

Transparency is an important issue for **VTO** algorithms too. These algorithms generate images of people wearing different clothes, and it is important for users to understand how these images are generated. One way to increase transparency is to provide users with information about the data used for training algorithms and the methods used for generating images. Another way to increase transparency is to provide users with more control over the generated images. This could involve allowing users to choose from a variety of different models and adjust the parameters of the algorithms.

3. Explainability:

The Implicit Trust and User Understanding Zone of the **Ethical 8 Risk Zone Framework** emphasizes the potential for AI to be used to manipulate and exploit users who may not fully understand how it works.

IBM AI Explainability Toolkits provide various explainability methods. It offers a wide range of explainer algorithms, including model-agnostic techniques like LIME and SHAP, as well as model-specific explainers for popular models like XGBoost and TensorFlow. Further, it supports different data types, which can be used to explain models that are trained on various data types, such as tabular data, text data, and image data. Moreover, an extensible architecture allows developers to easily add new explainability methods and integrate them with existing workflows. This enables the AI model to provide explanations for predictions that are understandable to humans.

The **Google AI Fairness Toolkit** has comprehensive documentation and tutorials that provide easy-to-understand instructions and examples for using the toolkit effectively. Plus, it improvises model evaluation and interpretation through tools for analyzing the fairness of trained models, which include explainable AI techniques to understand how models make predictions.

One of the challenges of **virtual try-on** algorithms is that they can be difficult to explain as they are often based on complex mathematical models that are not easily understood. One approach to increasing explainability is to use simpler models that are easier to understand. Another approach is to provide users with more information about the results of the algorithms. For example, users could be given information about the factors that the algorithms considered when making their predictions.

4. Robustness:

While the **Ethical 8 Risk Zone Framework** doesn't directly address robustness, its focus on fairness, transparency, and other key principles lays the groundwork for building robust AI systems.

One of the 5 key principles of the **IBM AI Explainability Toolkits** is robustness. It highlights that AI models should be able to function correctly in the real world, even when exposed to unexpected data or situations. Thus, explainability tools are structured to identify potential vulnerabilities in AI models and make them more robust. For example, they can help identify foolish model behaviour because of conflicting examples. By understanding how models make predictions, developers can identify and address biases, errors, and other issues.

The **Google AI Fairness Toolkit** tackles robustness by empowering developers to build models resilient against unforeseen data or situations through:

- Data exploration: Identify biases and potential vulnerabilities in training data, preventing skewed predictions.
- Counterfactual explanations: Understand how model decisions change with slight data modifications, pinpointing potential robustness issues.
- Robustness-aware training: Train models resilient to adversarial attacks and data inconsistencies, minimizing biased or unfair outputs.

- Model monitoring: Continuously track model performance under varied conditions, ensuring sustained fairness and reliability in real-world complexities.

The robustness of **virtual try-on** algorithms needs improvements too. For example, these algorithms may not work well on people with different skin tones or body types. This can lead to poor or no results for people who are not represented in the training data for the algorithms. They may not be able to get a realistic representation of how clothes would look on them. To make these more robust, the data should include people of all skin tones, body types, and sizes. Plus, algorithms must be designed to be robust to noise and outliers. This means that the algorithms should not be affected by small changes in the input data and fooled by data that is not representative of the population.

5. Privacy:

The Surveillance State, Data Control, and Monetization Zone of the **Ethical 8 Risk Zone Framework** embarks on the potential for AI to be used to create a surveillance state where our every move is monitored, and our data is collected and monetized without our consent. Adding to this, addiction and the dopamine economy zone concentrate on the potential for AI to be used to create addictive products and services that exploit vulnerabilities and manipulate behaviour. This resonates with the need for making models capable of handling user information correctly with due consent.

Moreover, one of the 5 key principles of the **IBM AI Explainability Toolkits** is value alignment. AI models should be aligned with human values such as justice and privacy. Explainability tools can help to ensure this by providing insights into how they make decisions and allowing humans to control the behaviour of models. And this personalization enhances privacy and trust among users.

The **Google AI Fairness Toolkit** respects user privacy in multiple ways, like:

- Focus on data anonymization and aggregation: It encourages anonymizing datasets used for training models, minimizing individual identification risks.
- Differentiated privacy analysis: Tools allow analyzing fairness metrics without revealing sensitive individual data, protecting privacy while ensuring fairness assessments.
- Support for privacy-preserving algorithms: It offers tools for training models that minimize data collection and storage, reducing privacy risks.
- Compliance with regulations: The toolkit aligns with privacy regulations like GDPR, ensuring responsible data handling and user control.

There are serious concerns about the privacy implications of **virtual try-on** algorithms. For example, these algorithms may collect data about users' bodies and faces, which could be used to identify them or track their movements. This data could also be used to create personalized advertising or to target users with malicious content. Thus, it's important to make sure that users are aware of what data is collected and how is it used, along with using strong security measures to protect it from unauthorized access.

Conclusion

By addressing ethical considerations and implementing the proposed framework, VTO can become a powerful tool for the fashion industry, promoting inclusivity, convenience, and

responsible technological development. The framework highlights the importance of transparency, explainability, and accountability in AI, encouraging developers and businesses to prioritize ethical values and ensure the responsible use of AI technology. This research contributes to the ongoing discussion on ethical AI development and provides a practical framework for its implementation in the fashion industry, paving the way for a more responsible and inclusive future for technology and fashion.

Citations

- McKinsey & Company. "Generative AI in Fashion." McKinsey & Company, 2023.
- Institute for the Future. "Playbook for Ethical Tech Governance." IFTF, 2023.
- IBM Research. "AI Explainability 360." IBM, 2023.
- IBM Research. "Introducing AI Explainability 360." IBM Research Blog, 9 Feb. 2021.
- Google. "Playing with AI Fairness." PAIR, 2023.
- Chouldechova, A. and A. Roth. "Fairness Tutorial." Google Sites, 2023.