**Ethics of explanation - Assignment Guidelines**

The purpose of this assignment is to apply the insights on the ethics of explainability to your own project.

|  |
| --- |
| **1. Introduction**  **Student names: Nir Tevet, Gal Granot**  **Project Title: Email AI Assistant**  **Describe your project**, i.e. what does the project do? what does it aim to achieve? (2-3 sentences)  The project aims to provide general assistance for email handling – ie classify emails as phishing or legitimate and direct users at important passages in legitimate emails. We used the attention architecture with BERT pretrained model. |
| **2. Have a large language model (LLM) answer the following questions on your project:**  2a. List 3 types of stakeholders that will be affected by the project.  2b. What will an explanation that is given to each stakeholder look like? (1 paragraph max)  2c. Who is responsible for giving the explanation to each stakeholder? (1 paragraph max)  **2a. List 3 types of stakeholders that will be affected by the project.**   1. **End Users (Email Users)**: Individuals who use email services and will directly interact with the classification system. 2. **IT and Security Teams**: Professionals responsible for managing cybersecurity within organizations, who will rely on the system to mitigate phishing threats. 3. **Email Service Providers**: Companies that provide email platforms and will integrate or host the system.   **2b. Explanation to Each Stakeholder**   * **End Users**: "This system will help you by automatically identifying and marking phishing emails, reducing your risk of falling victim to scams. For legitimate emails, it will highlight key parts, saving you time by directing your attention to the most important information." * **IT and Security Teams**: "The model will act as an additional layer of defense by filtering out phishing attempts before they reach your users, helping to reduce security incidents. It also provides contextual insights by highlighting crucial content in legitimate emails, enhancing user productivity." * **Email Service Providers**: "Our system integrates seamlessly with your platform, offering your customers enhanced security by detecting phishing emails and improving their email experience by spotlighting important information in legitimate emails."   **2c. Responsibility for Giving the Explanation**   * **End Users**: The customer support or communication teams of the email service providers will be responsible for explaining the system to the end users. They are well-positioned to convey the benefits and usage instructions in a clear, user-friendly manner. * **IT and Security Teams**: The organization's cybersecurity or IT management teams will take responsibility for communicating how the system integrates with existing security measures. They are equipped to discuss the technical benefits and implementation details. * **Email Service Providers**: The development or product management team of the project will communicate the details to email service providers, focusing on integration capabilities, system benefits, and potential impacts on service quality. |
| **3. Reflect on the AI output (this section requires your independent, creative thinking. Do not use LLMs for this stage!):**  3a. What do you think needs to be added/changed in the Generative AI responses, to make the explanation more ethical?  (1 paragraph max)  As an added consideration, we should make sure the model is secure and private – it accesses private information (all the organization’s users’ emails), and therefore there is a chance for a privacy leak. We should make the model does not capture actual text – perhaps by sending it hashed tokenized data and unhashing it with a separate hash function, so no one who controls the model can access the text underneath. |

List of LLMs to use (choose one or more to use while filling the template):

* ChatGPT: <https://chat.openai.com/>
* Claude: <https://claude.ai/>
* Gemini: <https://gemini.google.com/>
* You: <https://you.com/>
* Mistral: <https://chat.mistral.ai/>