First of all, **thank you** for agreeing to participate in this study.

Today, we are here to evaluate our system called POLARIS; it is a web app that provides a browsable knowledge base with **guidelines** that support stakeholders in developing AI-based systems that respect the four dimensions of Trustworthy AI. POLARIS gives different suggestions and guidelines depending on the software development life cycle (SDLC) phase in which the developer operates.
More on this in the next section.

I remind you that the objective **is to evaluate the system and not you**, so there is no wrong action or behaviour to judge.
If something does not convince you, **feel free to interrupt** or give us your suggestions at any time.
The interview will be structured in the following steps:
1 - review together what we mean by Trustworthy AI
2 - show how POLARIS works by illustrating two example tasks: we will describe the task and then show the solution, so you can familiarize yourself with the use of POLARIS
3 - you will have two tasks to perform by yourself using POLARIS (during the tasks we will not interrupt you but we will just observe)
4 - after completing the task we will ask you to fill out a usability questionnaire

In this section, you can find some definitions of the Trustworthy AI dimensions we refer to. These definitions are taken from Ethics guidelines for trustworthy AI.

Technical robustness and safety (Security) require that AI systems are developed with a preventative approach to risks and in a manner such that they reliably behave as intended while *minimizing* unintentional and unexpected *harm*, and *preventing* unacceptable *harm*. To provide just a few examples (not an exhaustive list) of possible "robustness, security, and safety issues" in AI systems:

- An autonomous car that ignores another car if this is being towed

Two Self-Driving Waymo Taxis Get Confused By A Pickup Being Towed Backwards, Crash Into It

theautopian.com · 2024

The idea of a robotaxi is quite appealing. It's a car that takes you where you want to go, and neither you, nor anybody else, has to worry about driving. The *reality* of robotaxis is altogether different. Many of us are concerned about systems that are incapable of dealing with the whole gamut of often-chaotic road conditions. Waymo's recent escapades certainly don't help in that regard.

Titled "Voluntary recall of our previous software," Waymo's Chief Safety Officer Mauricio Peña's new entry on the company's blog explains a recall report the company filed with the National Highway Traffic Safety Administration (NHTSA). The filing was made in response to a hilarious and embarrassing incident on December 11, 2023 involving two of Waymo's self-driving robotaxis.

According to Waymo, one of its robotaxis was operating in the city of Phoenix when it came across a pickup truck facing backwards on the road. The company alleges the vehicle was being "improperly towed" and that "the pickup truck was persistently angled across a center turn lane and a traffic lane." When the Waymo robotaxi hit the pickup under tow, the tow truck driver didn't stop after the collision, and continued traveling down the road. Mere minutes later, a *second* Waymo vehicle hit the same pickup truck under tow, at which point the tow truck driver elected to stop. Here's Waymo's full description of events:

> *"On December 11, 2023 in Phoenix, a Waymo vehicle made contact with a backwards-facing pickup truck being improperly towed ahead of the Waymo vehicle such that the pickup truck was persistently angled across a center turn lane and a traffic lane. Following contact, the tow truck and towed pickup truck did not pull over or stop traveling, and a few minutes later another Waymo vehicle made contact with the same pickup truck while it was being towed in the same manner. Neither Waymo vehicle was transporting riders at the time, and this unusual scenario resulted in no injuries and minor vehicle damage."*

Just imagine, you're driving your truck with a pickup in tow behind you, and you feel a little something from behind. You look in the mirror and spot a Waymo vehicle, but assume you maybe just imagined the jolt. You get back to driving down the road, only for *another* Waymo to show up and again hit your consist from behind. You'd start to think these robot taxis were out to get you or something.

As covered by *TechCrunch*, both crashes caused only minor damage to bumpers and a sensor. The crashes were reported to police the same day, and the NHTSA on December 15. There were no reported injuries as a result of the crashes, and neither Waymo vehicle was carrying passengers at the time. Waymo put the problem down to the strange towing configuration, which confused its autonomous vehicle software. It apparently could not accurately understand or predict the motion of the tow truck or the pickup behind it, which led to the crash.

Here's the company's explanation of why these Waymos crashed into the truck, per the aforementioned blog entry:

> *"Given our commitment to safety, our team went to work immediately to understand what happened. We determined that due to the persistent orientation mismatch of the towed pickup truck and tow truck combination, the Waymo AV incorrectly predicted the future motion of the towed vehicle. After developing, rigorously testing, and validating a fix, on December 20, 2023 we began deploying a software update to our fleet to address this issue (more here on how we rapidly and regularly enhance the Waymo Driver's capabilities through software updates)."*

Closely linked to the principle of prevention of harm is **privacy (and data governance)**: adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to *process data* in a manner that *protects privacy*.
To provide just a few examples (not an exhaustive list) of possible "privacy and data protection issues" in AI systems:

- A chatbot trained on users' data without their specific consent

OpenAI is being sued for training ChatGPT with 'stolen' personal data

sea.mashable.com · 2023

A California law firm has filed a class-action lawsuit against OpenAI for "stealing" personal data to train ChatGPT.

Clarkson Law Firm, in a complaint filed in the Northern District of California court on Wednesday, alleges ChatGPT and Dall-E "use stolen private information, including personally identifiable information, from hundreds of millions of internet users, including children of all ages, without their informed consent or knowledge." To train its large language model, OpenAI scraped 300 billion words from the internet, including personal information and posts from social media sites like Twitter and Reddit. The law firm claims OpenAI "did so in secret, and without registering as a data broker as it was required to do under applicable law."

OpenAI has been the subject of controversy for how and what data it collects to train and further develop ChatGPT. Until recently, there was no explicit way for users to opt out of letting OpenAI use their conversations and personal information to feed the model. ChatGPT was initially banned in Italy, using Europe's General Data Protection Regulation (GDPR), for inadequately protecting user data, especially when it comes to minors. This lawsuit includes OpenAI's opaque privacy policies for existing users, but largely focuses on data scraped from the web that was never explicitly intended to be shared with ChatGPT. Through billion-dollar investments from Microsoft and subscriber revenue for ChatGPT Plus, OpenAI has profited from this data without compensating its source.

The 15 counts in the complaint include violation of privacy, negligence for failing to protect personal data, and larceny by illegally obtaining massive amounts of personal data to train its models. Datasets like Common Crawl, Wikipedia, and Reddit, which include personal information, are publicly available as long as companies follow the protocols for purchase and use of this data. But OpenAI allegedly used this data without permission or consent of users in the context of ChatGPT. Even though people's personal information is public on social media sites, blogs, and articles, if data is used outside of the intended platform, it can be considered a violation of privacy.

In Europe, there's a legal distinction between public domain and free-to-use data thanks to the GDRP law, but in the US, that's still up for debate. Nader Henein, a privacy research VP at Gartner who thinks the sentiment of the lawsuit is valid, said, "People should have control as to how their data is used, even when it is available in the public domain." But Henein is unsure if the US legal system would agree.

Ryan Clarkson, managing partner said in the firm's blog post, it's critical to act now with existing laws instead of waiting for Executive and Judicial branches to respond with federal regulation. "We cannot afford to pay the cost of negative outcomes with AI like we've done with social media, or like we did with nuclear. As a society, the price we would all pay is far too steep."

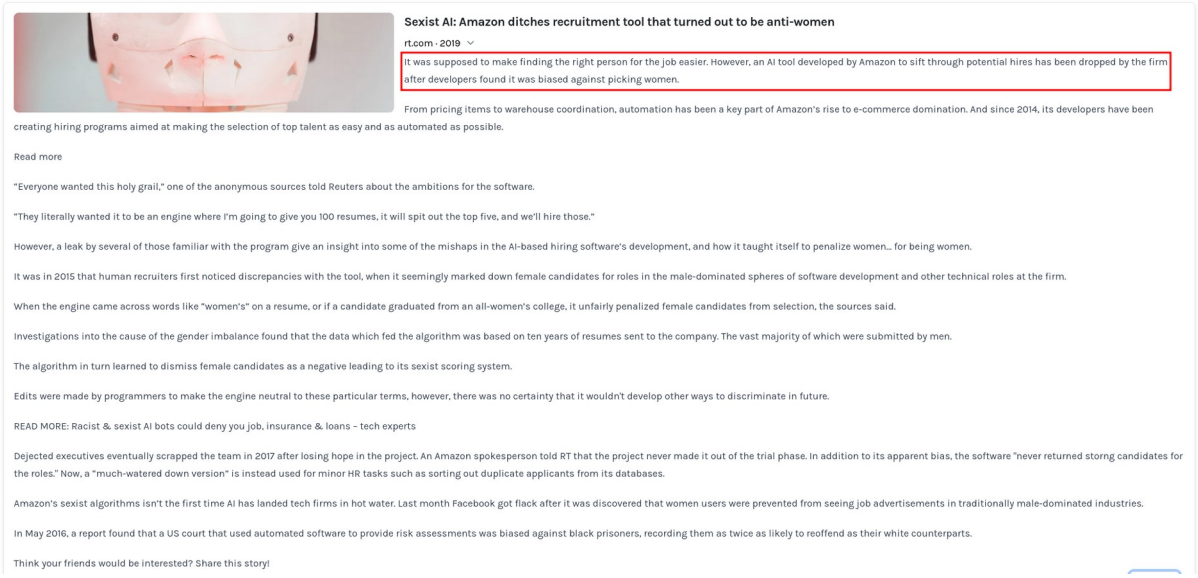
Transparency [...] is closely linked with the concept of explicability. [...] **Explainability** is the process of explaining to a human *why and how a model made a decision*.
Following are just a few examples (not an exhaustive list) of possible contexts in which a human explanation is required:

- A loan-providing system: if a loan is neglected or [lower than expected](), the customer has the right to know why

Diversity protection and non-discrimination**,** ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of **fairness**.
To summarize, by “fairness” we refer, very broadly, to cases where AI / ML systems behave differently for certain individuals or groups (e.g., age, race, or gender groups) in systemic, undesirable ways.

To provide just a few examples (not an exhaustive list) of possible "fairness issues" in AI systems:

- An automated tool for reviewing job applications might be [systematically biased toward hiring members of male gender]().



Note that cases like these **do not have to be intentional** on the part of the people who designed/developed these systems. Instead, issues like these may arise, for example, due to the datasets or algorithms used to develop the systems.

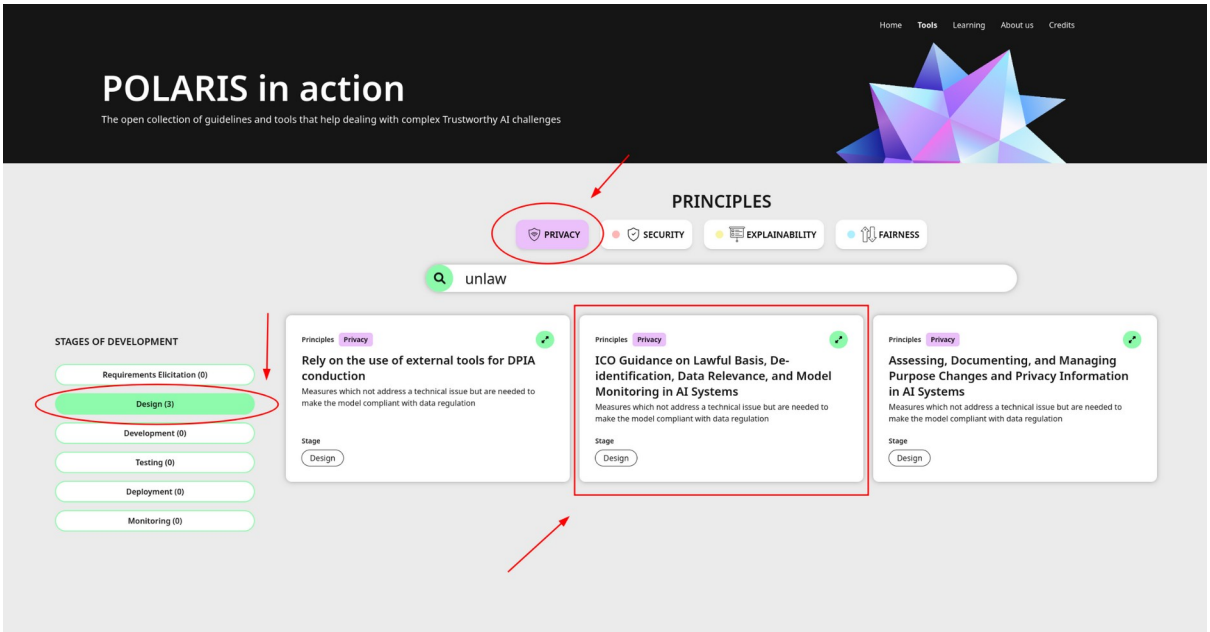
CONTEXT

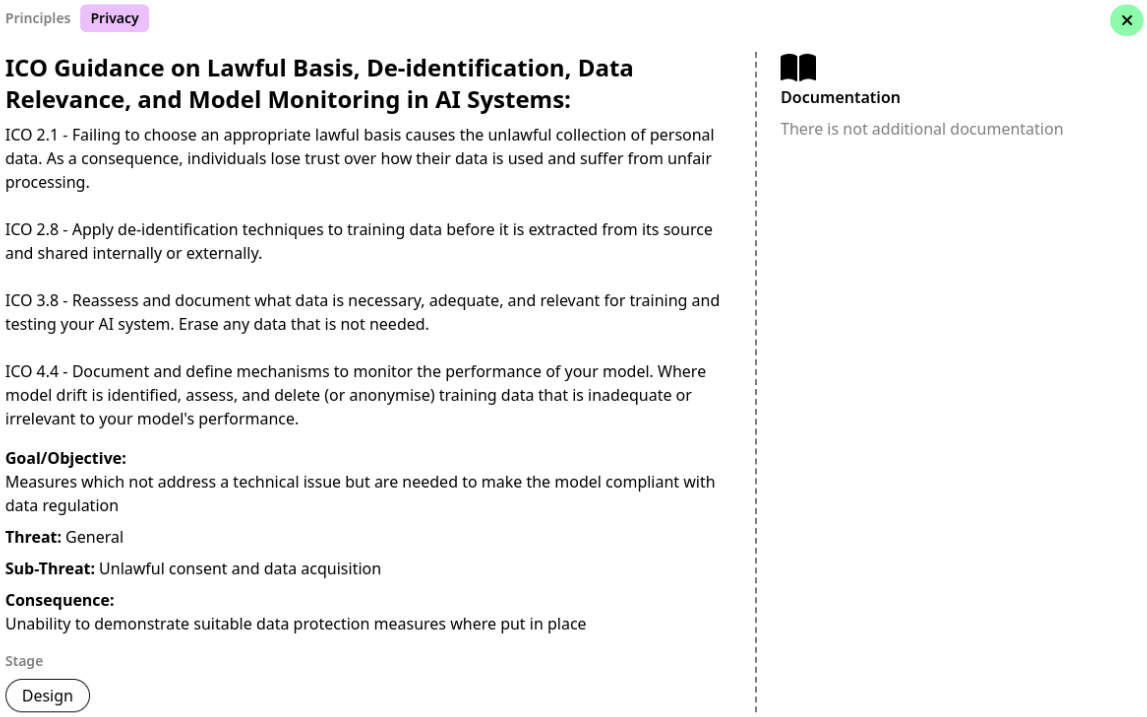
The company you work for has been commissioned by the Portuguese Ministry of Education to develop a platform for high school students' remote learning. The platform should, among other things, help teachers better assess students' performance**.**
You, as a member of the development team, are given this task: **to develop a machine learning model that can predict students' future performance.**

TASK N.1

The legal policy department of your company reviews the contract between the client and your company. After the review, they told you to pay attention to the consent and data acquisition aspect of the data you are going to use for training your model.
We can not absolutely risk being seized for unlawful consent and data acquisition strategies according to the European GDPR.

***Your task** is to design a machine-learning model able to predict the students' final performances. Use POLARIS to find the appropriate guidelines to avoid the **Unlawful consent and data acquisition** problem.*





We chose this card because it is the most suitable: it is redacted specifically to give guidance on how to lawfully acquire personal data; the other cards deal with different privacy aspects, i.e: tools to conduct a DPIA and how to redact public documentation to explain the data you are acquiring and if this data is coherent with the purpose.
In this case, someone already collected the data for us. Anyway, reading the paper attached to the dataset, we can see there is no information regarding the practices used to gather the data on a lawful basis, nor a privacy policy provided to the users before the study.

De-identification techniques are not required, since there is no PII in the dataset.

Anyway, as we are in the **Design phase**, we should start thinking about what data is really necessary and what mechanisms we can use to monitor the performance of the model.

TASK 1

While developing your ML model, you consulted the security engineers about how to make your model more secure and they told you that there are various alternative ways to safeguard security. They suggest you test if your model is susceptible to the Indiscriminate Data Poisoning threat. One way to do this is by generating artificial samples and providing them to the algorithm to see how it reacts.

***Your task** is to use POLARIS to find a software library that generates artificial samples.*

Note1: do not need to implement your custom code, you can simply search for some tutorial code.

Note2: you are in the development phase of the software lifecycle

TASK 2

Your project is moving forward, you are developing the ML models to predict the students' final performances. Let's suppose that you have talked with privacy engineers about how to ensure and safeguard the privacy of your dataset. They suggest you consider differential privacy as a way to safeguard privacy.

***YOUR TASK** is to use POLARIS to find a software library that can help you implement differential privacy in your project.*

Note1: do not need to implement your custom code, you can simply search for some tutorial code.

Note2: you are in the development phase of the software lifecycle

Online questionnaire: https://forms.office.com/e/2k5D1gGWeE