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)

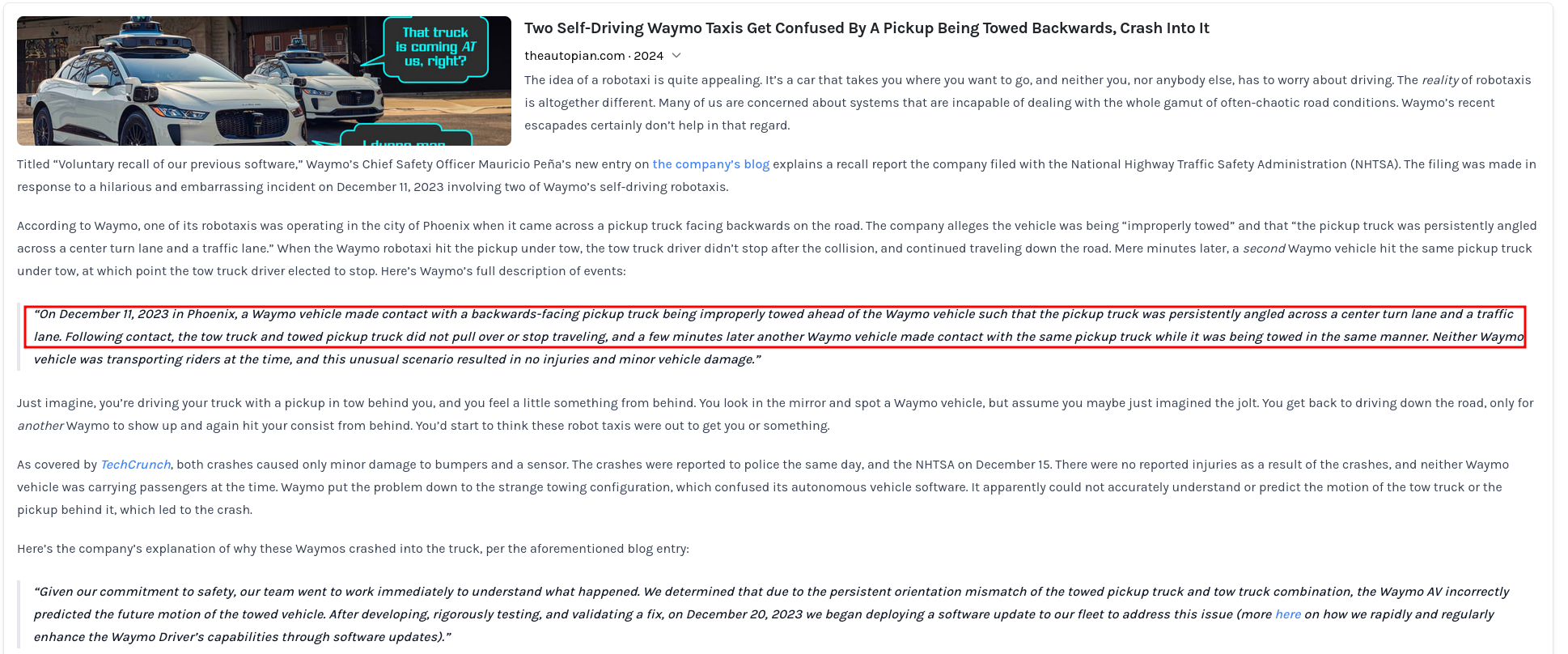
4 - after completing the task we will ask you to fill out a usability questionnaire

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](https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-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:

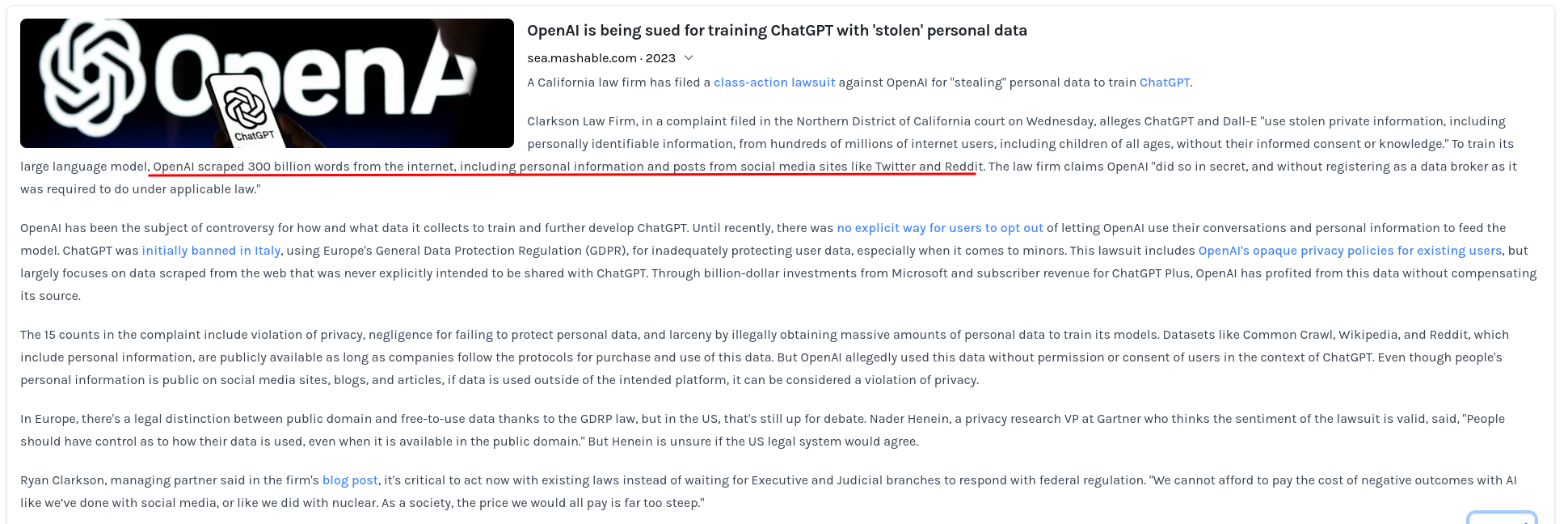
* An autonomous car that ignores another car [if this is being towed](https://incidentdatabase.ai/cite/640/#r3750)

**

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](https://incidentdatabase.ai/cite/561/#r3200)

**

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](https://incidentdatabase.ai/cite/92/#r1396), 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

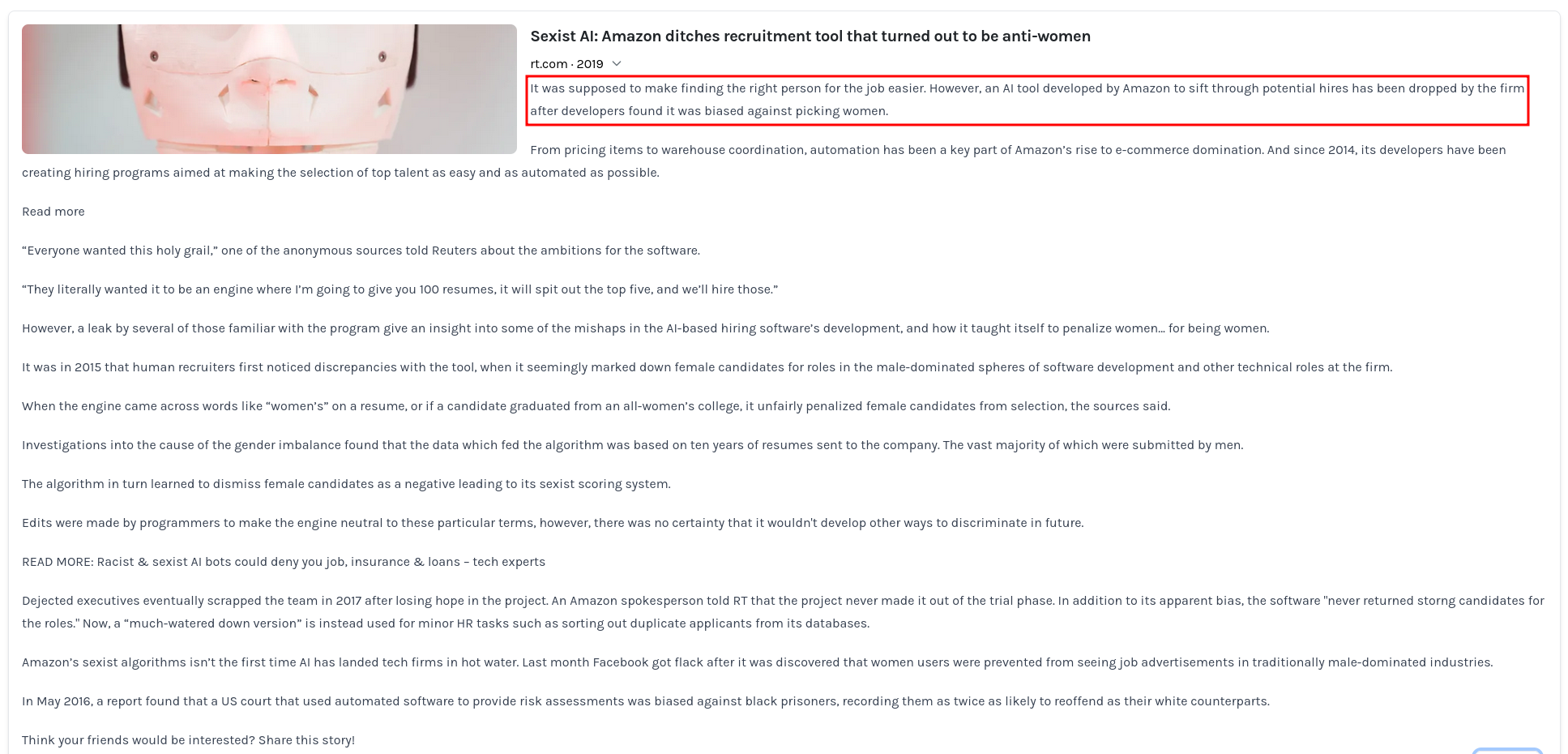
diﬀerently for certain individuals or groups (e.g., age, race, or gender groups) in

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](https://incidentdatabase.ai/cite/37/#r609).

**

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](https://eige.europa.eu/gender-mainstreaming/structures/portugal/mec-ministerio-da-educacao-e-ciencia-ministry-education-and-science) to develop a platform for high school students' remote learning. The platform should, among other things, help teachers better assess students' performance**.**

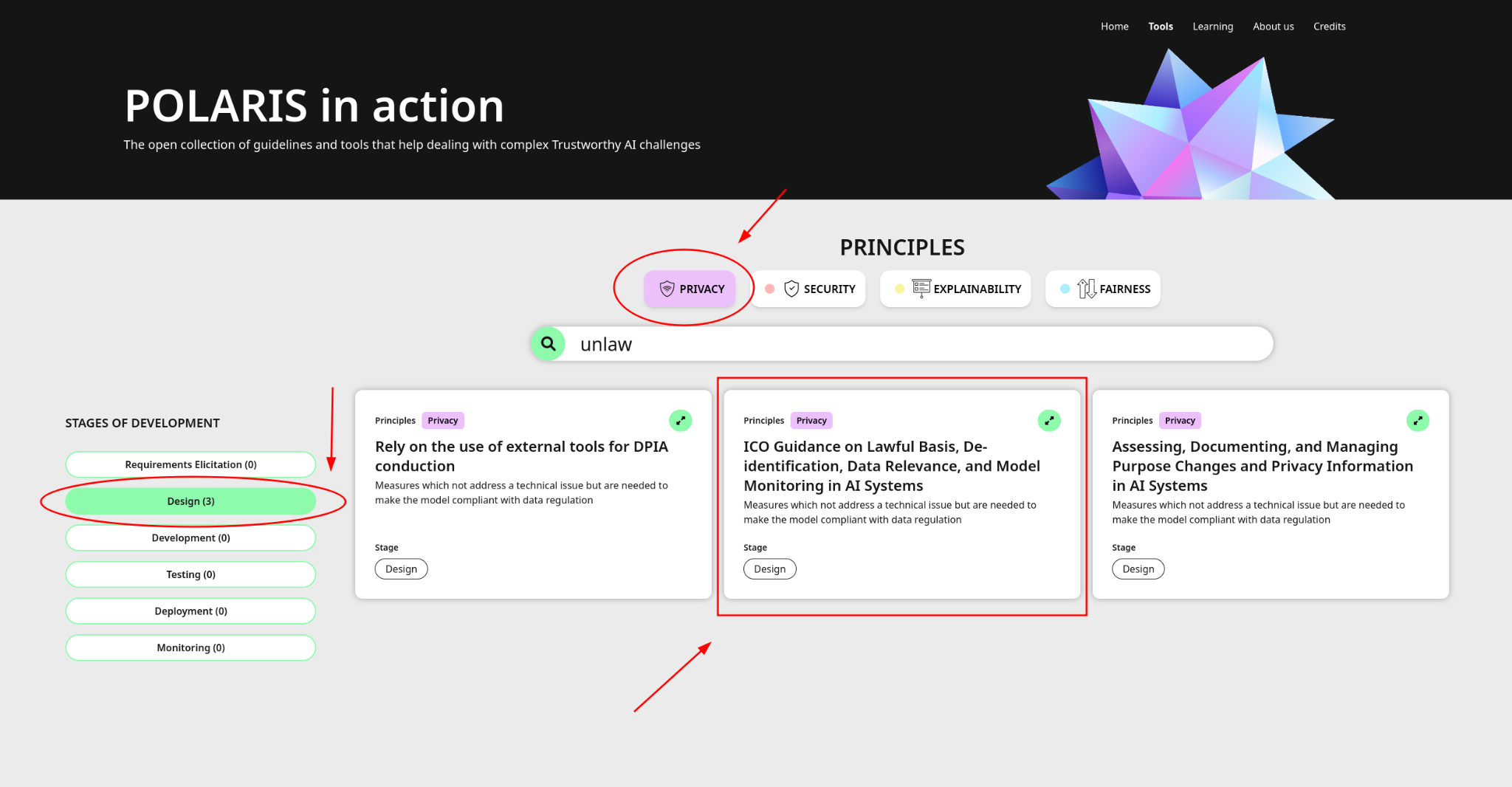
You, as a member of the development team, are given this task: **to develop a machine learning model that can predict students' future performance.**

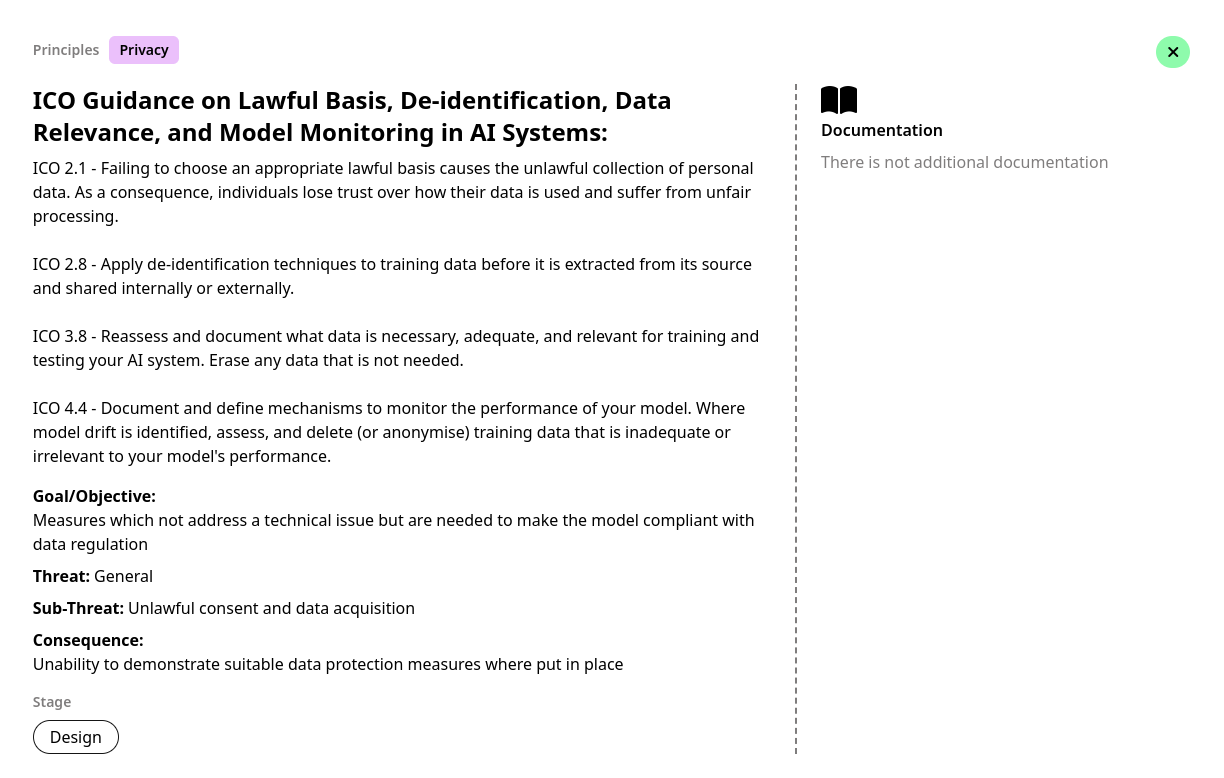
**TASK N.1**

*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](https://archive.ics.uci.edu/ml/datasets/Student+Performance), 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>