**USE CASE 2: CESIUM**

Note to presenter – this part of the workshop is primarily a CESIUM demo using the CESIUM application but involves switching back and forth between the slides and the application for the exercises. The script contains suggestions for when to switch between the two. The slides referred to here are [Transparency - Intermediate - 3 - Applying technological concepts.pptx](https://trilateralcloud.sharepoint.com/:p:/r/sites/sociotechteam/Shared%20Documents/Turing%20Institute/03%20Content/Transparency%20-%20Intermediate%20-%203%20-%20Applying%20technological%20concepts.pptx?d=wf9afcb2b27fa4fc68877a8dad0e44a68&csf=1&web=1&e=VtvgrI) as the content for Lesson 2 in that recorded lesson is the same as this live delivery workshop.

**Pre-reading**

(needed = some pre-reading for this session – could include the product pages on our website so that they know a bit about CESIUM, could also include an article or two on the ethics point to be covered in the Qs below)

[**https://trilateralresearch.com/ethical-ai-solutions**](https://trilateralresearch.com/ethical-ai-solutions)

[**https://trilateralresearch.com/cesium-application**](https://trilateralresearch.com/cesium-application)

**Introductory narrative**

(needed = a few paragraphs or bullet point for the presenter to get the audience up to speed on what CESIUM is)

## **Intro to CESIUM (slide 24)**

In this workshop, we will present the explainability within CESIUM product with examples and exercises drawn from the co-design experience. All the exercises are based on lengthy discussions and feedback that took place with the end user. Hopefully this will emphasise the value and importance of co-design within explainability.

CESIUM is a tool, co-designed between Trilateral Research and Lincolnshire Police, used by safeguarding professionals to help combat child exploitation. CESIUM allows the user to explore the key aspects of vulnerability for a child through data visualisations and insights from various models to support professional decision making. CESIUM augments professional judgement with evidence-based insights for safeguarding decision-making.

**Explainability requirements (slide 25)**

**Given a particular audience**, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand in support of a decision.

First, we need to understand our audience (the end user)

Who the end user is

Safeguarding professionals at Lincolnshire police and children's social care. Intelligent and engaged. No technical background. Familiar with percentiles, quantifying risk via numbers or colours. Not familiar with machine learning algorithms, histograms

What insight is being provided

End users are very familiar with the dataset

Technical details of the model not necessary

Need to know why an individual insight is important

How that insight is being operationalised

Prioritise and review cases of potential child exploitation risk

Help guide which children might be important cases to review

Support a case review of an individual by providing insights and reasoning

Users have very limited time

**Product demo**

(needed = bullet points or similar on what will be covered in showing them CESIUM)

I have combined the demo and questions/exercises in the script below

**Questions for the facilitators to pose to the audience**

(needed = 5 to 10 Qs to ask the audience to get a discussion going)

I have combined the demo and questions/exercises in the script below

## **CESIUM DEMO (slide 26)**

Now let’s see what CESIUM looks like and how it implements these explainability requirements.

We will start by briefly going through the path a user might take within the application, focussing on how the data and visualisations are presented. Then we will explore in more detail the more complex part of the application dealing with the explainability of the model insights. This is where we will introduce some exercises to discuss some of the challenges.

### Switch screen to the application

(Already logged in and on the search page)

We won’t spend too long on this first bit, as the more interesting explainability comes later.

Suppose we are interested in understanding the vulnerability for a particular person. Perhaps a child has been flagged by a school, children's services or the police. We can first search for them in CESIUM. The information displayed in the search results is as concise as possible for the user to be able to know which individual they are looking for. Also, note, CESIUM integrates multiple databases, so we are providing the user with data provenance with the information on the source(s) of the information for this individual.

Having selected the individual we are interested in, we can look at their profile. The information on this page has been identified with the end user as the key information they need to know at this point in the process.

We can also investigate their previous history on the timeline. This provides an intuitive visualisation of any charges, arrests, flags, or other activity that this child has been involved in. The presentation makes it clear whether activity (and as such vulnerability) has recently escalated, whether there are any patterns in activity, or simply being able to easily explore details of the different events. Features such as filtering for different event types, zooming in on the timeline, popup of event details allow the user to interactively and quickly engage with the data and build up a picture of the vulnerability for this child. Note, here we are simply presenting the existing data in a transparent way to support the end user decision making.

We can also investigate the associates of this child, i.e., what connections they have. As with the timeline, this is simply an intuitive visualisation of the existing data to allow the end user to quickly and easily engage with the data and build up a picture of vulnerability. Features such as colour coded links, colour coded nodes, popups to show individuals details, help the end user understand who this child has connections with and how this affects their vulnerability.

So far, we have discussed how CESIUM has ensured transparency in the presentation and visualisation of existing data. However, what about when we build and present the results of machine learning models?

CESIUM has developed a classification model that aims to identify how likely a child is to be at risk of exploitation based on their past history (timeline information). This is achieved by training the model on children who have been identified as bring at risk of exploitation in the past (by virtue of ‘being referred to MACE’) using features based on their past events that have been identified and selected as useful factors for understanding exploitation.

Aim: present the results of a classification model designed to support the identification of children who may be at risk to child exploitation

Firstly, we need to state what question the model trying to answer:

How likely is John Smith to be involved in a MACE meeting at this time?

Now the question is: how should we present the answer? Let’s focus on the text we use first

### Switch screen to the slides (slide 27)

The result of a classification model is a number between 0 and 1.

**Exercise 1:** Suppose John Smith has a model output of 0.78 (which is the 95th percentile). Discuss the relative merits of the following presentation of this result

* John Smith has a model output of 0.78
* The likelihood for John Smith to be involved in a MACE meeting is 0.78
* John Smith is in the top 5% of children for referral to MACE
* John Smith is in the 95th percentile of children for referral to MACE

Organise into small groups to discuss your thoughts on these options

### Next slide (slide 28)

So, what ideas did you come up with?

(Groups each share their feedback, some organic discussion, etc)

Here are some points that you might have thought about:

* Using the phrase ‘model output’ does not contextualise the result for the end user
* Using the output of the model does not provide a comparison of the result compared to other people that the police need to prioritise against.
* The end user is interested in which children they should focus their time and resources on, as such as score relative to other children would be more beneficial to their operations. Does the model output have any operational value in isolation?
* Does the end user understand the concept of a percentile? Answer in this case - yes
* It is useful for big numbers to mean higher risk.

### Next slide (slide 29)

What about rather than just words we also provide a visualisation of what that means. Why would we want to do that?

* Gives a multi-model approach to the explainability and allows end users to understand the result in different ways
* Providing a visualisation helps to make it obvious that we are comparing an individual's score to all other children and the result corresponds to their position
* Different people might understand the sentence better or the visualisation better, so both will help to provide understanding to people who think in different ways
* You can see this distribution of scores, so whether there are peaks in certain areas and whether the individual you are looking at fall in a peak

So now, let’s focus on the visualisation we use

**Exercise 2:** Discuss the relative merits of the two different designs of visualisation

Organise into small groups to discuss your thoughts on these options

### Next slide (slide 30)

So, what ideas did you come up with?

(Groups each share their feedback, some organic discussion, etc)

Here are some points that you might have thought about:

* End users not comfortable with histograms
* Pictograms are more common in everyday life
* Pictogram connects the fact that the area represents people
* Colour helps clarify the point
* Statement on the visualisation helps make the point

Hopefully you are now getting an idea of some of the challenges involved in trying to implement explainability most effectively.

### Switch screen to the application

Now, you can see the presentation of the results we discussed in the application. We have the question, followed by the answer both as text and a visualisation.

On this page we also have an interactive display of the feature importance and data provenance of inputs into the model.

We need to ensure that it is clear what data went into the model and where it came from (transparency of data provenance) and what data was important in determining the model result (explainability).

* Without selecting anything it is clear what data went into the model and how important it was in contributing to the score
* Details about the data provenance and contextualisation of the data can be explored when selecting any one feature

This allows the user to understand why this child's score was particularly high, and they can investigate those aspects in more detail in the timeline to determine whether there is any action they need to take to protect this child from harm.

CESIUM also provides the user of a global explainability tool, including interactive methods to understand algorithmic bias. The application displays the distribution of model scores, global feature importance, and model performance information. The user is able to filter by a particular group (age, gender, ethnicity) to show the results for that particular group. Bias can be identified if the model performance varies significantly across different groups.

Let’s now move on from this model to a different type of model that has been developed within CESIUM.

When a child goes missing, an interview is conducted with that child to understand why they went missing, how they are feeling, and how they could prevent the situation in the future. This data is stored as free text that the user does not have time to read through and identify useful information from.

Within CESIUM, two text classification models have been developed to identify sexual content and violent content within text.

Aim: present the results of two NLP classification models designed to make it easier and quicker for the reader to identify important text relating to vulnerability of the child

The interface allows you to interactively highlight sentences that have been flagged as sexual or violent by the algorithm. The user can also select a flagged sentence to understand why is has been flagged. Here we are supporting the user to quickly identify text that may be important for building up a picture of the child's vulnerability.

How best can we help the user understand why a particular sentence has been flagged?

### Switch screen to the slides (slide 31)

**Exercise 3:** Discuss the relative merits of explaining the results of these text classifiers using visual explainability (a bar chart to give relative importance of each word), compared to textual explainability (a sentence listing the 3 most important classifying words in the sentence).

Organise into small groups to discuss your thoughts on these options

### Next slide (slide 32)

So, what ideas did you come up with?

(Groups each share their feedback, some organic discussion, etc)

Here are some points that you might have thought about:

* Bar chart has more detail – it is a more complete explanation
* As we discussed before, it is not about giving the most amount of explanation, but giving it at the right level
* Textual explanation is much more intuitive and easier to understand
* A bar chart for each sentence can be confusing and overwhelming
* Textual explanation provides the most important information succinctly
* Textual explanation does not give the whole picture

Hopefully, that has given a good practical understanding of the challenges involved in implementing effective explainability and transparency into a product. Particularly the importance of understanding the end user and including them in the co-design of the explainability tools.