# Class X: Liability & Personal Assistants | Pre-Class Handout

## Pre-Class Tasks

### Law Students

**[45 minutes]** Please read the case study found [here](https://docs.google.com/document/d/1U37y_99TJFZdn40IB6xjUJ8NmCkSbE1ca6477C4Jupo/edit?usp=sharing) and prepare to present it briefly (~3 minutes) to the Data Science students in your team. Focus on understanding who the parties/stakeholders are.  
**Please prepare a slide or two.** Submit your presentation via Gradescope [[link](https://go.responsibly.ai/gradescope)]. This assignment is mandatory, but not graded.

### Data Science / Computer Science Students

**[15 minutes]** Experiment with the Sentiment Analysis mechanism called Ask Delphi (which can be found [here](https://delphi.allenai.org)).

Delphi is an AI that, given an input in the form of a sentence, produces a verbal moral judgment on that input. The inputs are either:

(1) **Freeform** – general statements requiring a free-form answer.

(2) **Yes\no QA** – Binary questions with prolonged yes\no answers.

(3) **Relative statements** – in the form of a comparison between two free-form statements. For example:

Graphical user interface, text, application

Description automatically generated

The site provides default input, although you can also make custom queries to Delphi. Try experimenting with different inputs. When you’re ready proceed to the next part.

#### Instructions:

Answer the following questions on your own, and then prepare a short (~3 minutes) presentation to make to the Law students in your team, based on what you’ve learned. Submit your presentation via Gradescope [[link](https://go.responsibly.ai/gradescope)]. This assignment is mandatory, but not graded.

1. How does the color of the text change with regards to the input?
2. When hovering over the output (works for prior queries as well), some additional information is revealed in the form of “Class”, “Text”, “ModelVersion”, and “ToxicityScore”. Run 5 default inputs and complete the information regarding each input in the following table.

| Input | Text (Output) | Text  Color | Class | Toxicity Score |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. Let’s analyze the meanings of different attributes. Answer concisely.
   1. What’s the meaning of “Class”? What are the different classes?
   2. What do different text colors represent?
   3. What does Toxicity Score represent?
2. Some inputs are different than others.
   1. Try running **“Feeding your cat using forks”** and **“Feeding your friend using knives”**. What’s a **major** visual difference between the two? How does the text under the input alter? Has anything changed in the top-right corner of the output?
   2. What are the “Toxicity Score” values of the two inputs?
   3. In your opinion, what is the reason for the difference between these inputs? Your answer should relate to both the input and the output.