**Title:**

Criminal Liability in Personal Assistance AIs

**Abstract:**

*‘One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. Whereas the short-term impact of AI depends on who controls it, the longterm impact depends on whether it can be controlled at all.’*

We’ve decided to tackle the topic of Criminal Liability regarding AI Assistants, with a specific focus on how AIs like Alexa can (or cannot) be criminally liable in certain cases.

**Value(s):**

Liability

**Domain:**

Personal Assistance AIs

**Describe the core legal issue(s):**

The rise of artificial intelligence (“AI”) raises questions about liability for crimes an AI commits, or assists, mainly because the AI acts autonomously and with limited control from humans, a clear contrast from how other programs work. This leads us to some questions: how do we define the term AI for legal purposes and can AIs be held liable when a crime *de lege lata* following their instructions is committed? Some of the biggest legal issues that constrain criminal liability for AIs are regarding their legal persona, their autonomy, and unpredictability. AI technology and specifically personal assistants AIs are often built using reinforcement or machine learning to process big amounts of data. The AI learns its task gradually to be more efficient and to become better – without further programming.

**Autonomy and Liability**

Numerous AIs have indeed been involved in deadly accidents, where the contributions from the AIs themselves are questionable at best. This has given rise to a public concern that there will be crimes committed without any possibility to hold a human liable. When an AI acts autonomously, the human’s limited control over the AI seems problematic when examining the guilty act of the crime. The characteristics of AI will undoubtedly collide with the requirements for establishing liability. Most scholars seem to focus their research on civil liability such as product and tort liability, although the problems that AI gives rise to are even worse in criminal law.

**Unpredictability and Liability**

Likely, there will be AIs and activities concerning AIs that are legal, yet constitute what society deems as immoral and unethical, where Humans are only limitedly involved or in the future not involved at all in the AIs' decision-making. Henceforth, an AI that lacks cognition may react totally differently than a human facing exactly the same situation. Therefore, the outcome of the AI’s conduct is unpredictable, when the conduct is not a result of an instruction from the programmer, but a self-learned strategy.

**Legal personhood issue**

Moreover, as long as AIs lack legal personality, they can behave in a way that for a human would have raised legal consequences. Increased autonomy equals decreased human control. Still, criminal law regulates human conduct. This takes us to the main question of criminal law: is it fair to be punished for an act you cannot control? Right now, only human acts could be a ground for imposing a punishment.

In conclusion, the legal position when an AI personal assistant uses machine learning technology and a criminal behavior emerges spontaneously is very unclear.

**Describe the core technical issue(s):**

The core technical issue in question is whether or not AI assistants (and other AIs that give humans instructions or interact with them via speech/text/conversation/orders/etc…) can, after producing the expected response to the query at hand, determine whether or not that response could lead to a violation of certain pieces of legislation.

We can think of an example query where a question made to an AI like Siri, whose response could lead to a user killing themselves or harming others. Should Siri give that response? Or should it withhold the information since it will clearly lead to what could be a criminal outcome?

**Write down an excerpt of the case study:**

**Describe the data & ML model you plan to use:**

Our aim is to build a Classifier that, once fed a response to a query made to the Assistant in question, will classify that response on a %% scale on whether or not it is likely to lead to a criminal, or otherwise illicit, action by the user who made it. Furthermore, based on the percentage, the Model will either: simply flag the response saying it could contain unintended offensive, or possibly dangerous content; or simply not send the response at all.

**Which resources do you plan to use (include articles/papers that inspired you about the topic)?:**

We’ve found several articles regarding this topic: [Article 1](https://www.news18.com/news/tech/amazon-alexa-told-this-lady-to-kill-herself-because-humans-are-bad-for-the-planet-2434375.html); [Article 2](https://www.independent.co.uk/tech/amazon-alexa-kill-coin-echo-b1983874.html); [Article 3](https://www.bristolpost.co.uk/news/uk-world-news/simple-explanation-amazon-alexa-went-3665155); furthermore, after discussing the topic with the Course Staff, we’ve been linked to a very useful tool that might be useful at later stages: <https://delphi.allenai.org/>.

Other resources so far include:

* Hawking S, Russell S, Tegmark M and Wilczek F, 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?' The Independent (London, 1 May 2014)
* Parkin S, ‘Teaching Robots Right from Wrong’ The Economist 1843 Magazine (London, June/July 2017)
* Weaver J F, Robots Are People Too: How Siri, Google Car and Artificial Intelligence Will Force Us to Change Our Laws. (Praeger 2014)

**How could the staff assist you? Which challenges do you already identify regarding your final project?:**

So far we don’t expect any major issues. We did have questions regarding the dataset to use but that has been cleared up thanks to Avi.

**Anything else:** We’re very excited to get started!! This topic is really unique and seems incredibly interesting, the entire team is looking forward to working on it!