# Week 5 - SGTA

**Santoni de Sio and Mecacci (2021) identify four responsibility gaps: a culpability gap, a moral accountability gap, a public accountability gap, and an active responsibility gap. Out of these, choose two gaps that you find most relevant and interesting. How can these gaps be characterised? Do you agree or disagree with Santoni de Sio and Mecacci's (2021) assessment of these gaps? Provide reasons for your evaluation.**

Santonio de Sio and Mecacci (2021) characterise the culpability gap in terms of "the problem of many hands" (Bovens, 1998; Thomson, 1980), in which culpability or blameworthiness is difficult to attribute due to the number of different parties responsible. They use the example of a driver using Driver Assist technology developed by company X, with digital systems developed by company Y, and ML algorithms developed by company Z and so on. This gap is exacerbated by the complexity of socio-technical systems, where multiple actors contribute to the design and operation of AI, leading to situations where no single individual can be fairly blamed. This ties into their characterization of the moral accountability gap, which describes the responsibility for individuals to answer "why-questions" pertaining to the reasoning behind certain actions. AI systems make this especially difficult by being opaque by nature, making it difficult to determine the exact reasoning as to why a conclusion was made or action taken.

I agree with their assessment, both gaps highlight significant ethical concerns regarding the deployment and utilization of AI in different real world scenarios. Their characterization of these gaps mirrors social and legal issues present today with the rise in self-driving cars and more AI-powered applications.

**Hindriks and Veluwenkamp (2023) deny that the employment of AI systems can create responsibility gaps. Instead, they argue that the problems identified by previous research are better described as control gaps. Reconstruct and critically discuss their argument. Is their argument valid and plausible?**

Hindriks and Veluwenkamp (2023) reject the concept of responsibility gaps, instead framing the issue as a control gap. They argue that machines should be designed to meet a morally acceptable level of risk, at which point the machine can be deemed not culpable in the event of an accident or wrongdoing. By reframing the argument, they propose that harm in this scenario is either blameless or can be indirectly attributed to 'enablers' such as designers and regulators. They frame this in terms of an unavoidable human driver accident, in which flaws in design or policy leaves the designers or policymakers culpable in the accident.

While this argument is valid in theory, it struggles to provide practical answers. As Santonio de Sio and Mecacci (2021) argue, when there are multiple developers contributing to a product traceability becomes practically difficult. This is also compounded by the opacity of AI systems in general, making it difficult to determine if there is a control gap present in the AI model's decision making. Additionally, a morally acceptable level of risk is entirely subjective and context dependent. While the argument frames the problem in a way that enables a technical solution, there are still issues to address before a implementable solution is found.