# Exploring Stakeholder Perspectives on Autonomous Systems & Answerability

THE UNIVERSITY *of* EDINBURGH

Louise Hatherall*, Dilara Keküllüoğlu, Nadin Kökciyan, Michael Rovatsos, Nayha Sethi*, Tillmann Vierkant, and Shannon Vallor (*based in CBSS, Usher Institute)

Advances in AI amplify longstanding concerns about ‘responsibility gaps’ that can undermine social trust in autonomous systems (AS) (Matthias 2004). Such gaps are thought to arise from a lack of human control and/or knowledge of, AS actions.

Recent literature proposes a shift from an agent-centered focus on explainable systems to a more relational, dialogical focus on ‘answerability’ to the ‘patients of responsibility’ (PoR) who are affected by, and vulnerable to, the acts of an AS (Coeckelbergh 2020).

Answerability is an approach to moral or legal responsibility (Duff 2009) that recognizes that humans build trust by ‘answering’ to one another for actions through explanations but *also* justifications, excuses, restitutions, apologies, reforms, promises, penalties etc.

It remains unclear, however, the extent to which answerability can be satisfied by sociotechnical systems with high degrees of machine autonomy, diffuse human agency, and no dialogical moral agency within the software or artifact itself (Tigard 2021).

Responsibility gaps are not a unique AS problem, as humans often don’t understand or control the causal etiologies of their behaviour. Yet humans routinely bridge such responsibility gaps through moral dialogue with other impacted parties, even without viable explanations. This offers a constructive way forward in bridging responsibility gaps in AS. We aim to show how answerability practices can promote the cultivation of responsibility for AS actions in our wider ‘moral ecology,’ even when the morally responsible agent(s) remains underspecified or uncertain.

This interdisciplinary project (drawing on expertise in philosophy, law, and informatics) proposes a novel approach to design for AS answerability, based on an instrumentalist framework of ‘responsible agency cultivation’ drawn from philosophy and cognitive science, as well as empirical results from structured interviews and focus groups in the application domains of health, finance and government. We outline a prototype ‘mediator’ agent informed by these emerging results, designed to help bridge structural gaps in organisations that impede responsible human agents from answering for AS.

## II. Exploring Stakeholder Perspectives

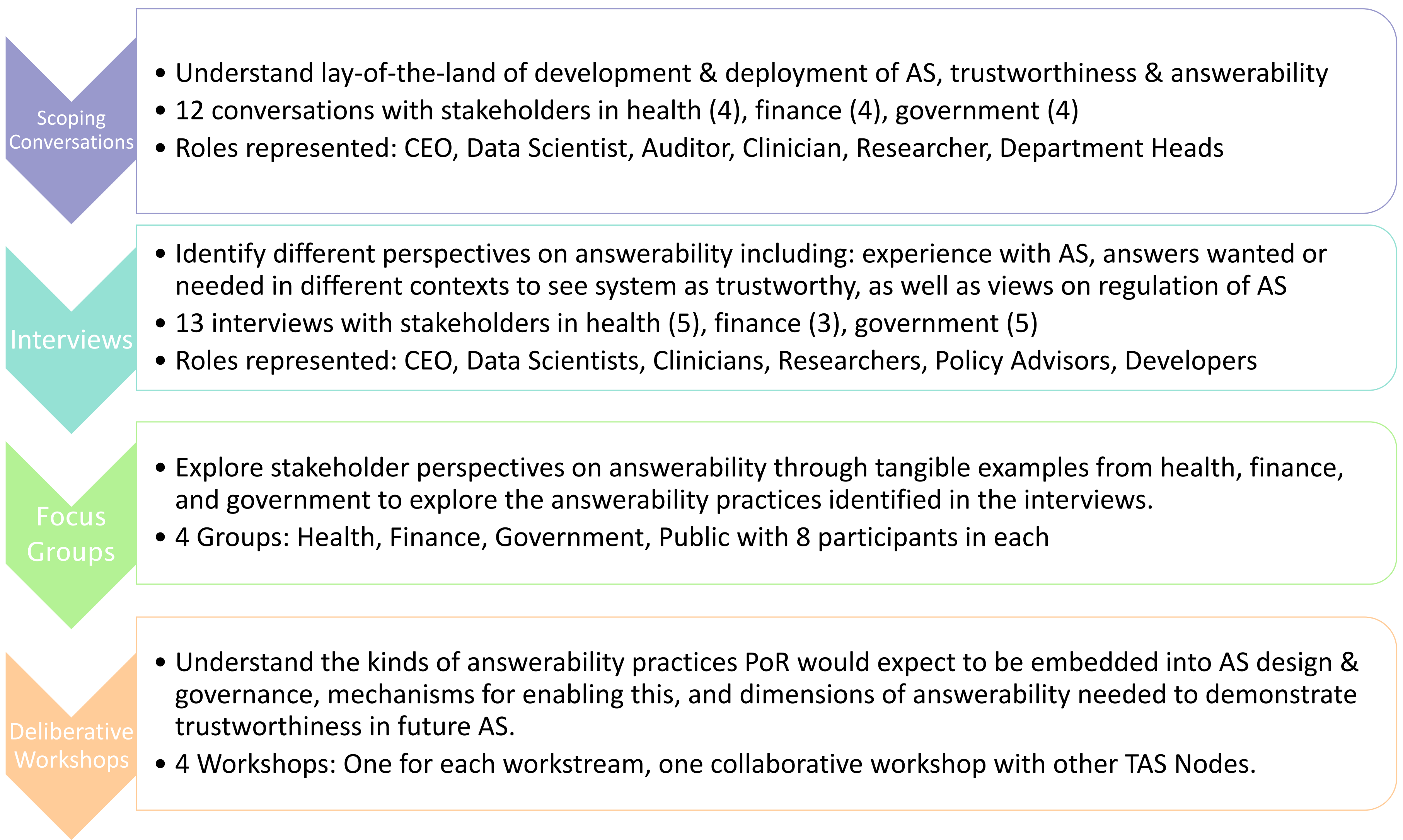
In order to align AS design, uses and regulatory efforts with the answerability needs and expectations of PoR, we need to better understand those needs and expectations. AI is increasingly impacting a range of domains; to remain manageable this research focussed on three areas where AI/AS are developing rapidly, one of which is health. Responsibility in AS and healthcare systems have attracted a range of research, and identified challenges to achieving trustworthy AS arising from uncertainty around decision making in the use of AS, concerns about bias leading to poorer health outcomes, and uncertainty around governance (Health Education England 2022; Richardson 2021). Important epistemic gaps remain around what PoR needs and expectations are, and how they vary for different vulnerable AS stakeholders and application contexts. This project aims to fill this gap by using socio-legal methods to:

- identify the types of answers that a range of stakeholders will need to trust AS in health, finance, and government;
- evaluate developing regulatory frameworks for their alignment with AS answerability;
- explore how to embed answerability practices within both regulatory and systems design, and assess the regulatory implications of doing so

UKRI Engineering and Physical Sciences Research Council

UKRI Trustworthy Autonomous Systems Hub

## Multi-pronged Empirical Data Collection



Scoping Conversations
- Understand lay-of-the-land of development & deployment of AS, trustworthiness & answerability
- 12 conversations with stakeholders in health (4), finance (4), government (4)
- Roles represented: CEO, Data Scientist, Auditor, Clinician, Researcher, Department Heads

Interviews
- Identify different perspectives on answerability including: experience with AS, answers wanted or needed in different contexts to see system as trustworthy, as well as views on regulation of AS
- 13 interviews with stakeholders in health (5), finance (3), government (5)
- Roles represented: CEO, Data Scientists, Clinicians, Researchers, Policy Advisors, Developers

Focus Groups
- Explore stakeholder perspectives on answerability through tangible examples from health, finance, and government to explore the answerability practices identified in the interviews.
- 4 Groups: Health, Finance, Government, Public with 8 participants in each

Deliberative Workshops
- Understand the kinds of answerability practices PoR would expect to be embedded into AS design & governance, mechanisms for enabling this, and dimensions of answerability needed to demonstrate trustworthiness in future AS.
- 4 Workshops: One for each workstream, one collaborative workshop with other TAS Nodes.

## Tentative Initial Findings

**Gaps in Healthcare:** There is empirical evidence of responsibility gaps in healthcare where AS are used, undermining the trustworthiness of such systems for both healthcare professionals and patients. Challenges also arise from the categorisation of roles in healthcare (Laurie 2017) leading to blurred regulatory distinctions (e.g. between those who use systems and those impacted by them).

**Answers & Assurances:** PoR want answers which are *practical* (how?), *reason based* (why?), and *relational* (who else?). Relational answers are more important in health (does my clinician use it often? Do other patients like me rely on it?). PoR may also want assurances about *their own* abilities interacting with the system.

**System Answers:** PoR will accept answers from systems in low-stakes contexts, as long as answers are: *accurate* (or clear about the level of accuracy possible), *dialogic* (particularly to initially probe knowledge or accuracy of the system), and *concise*.

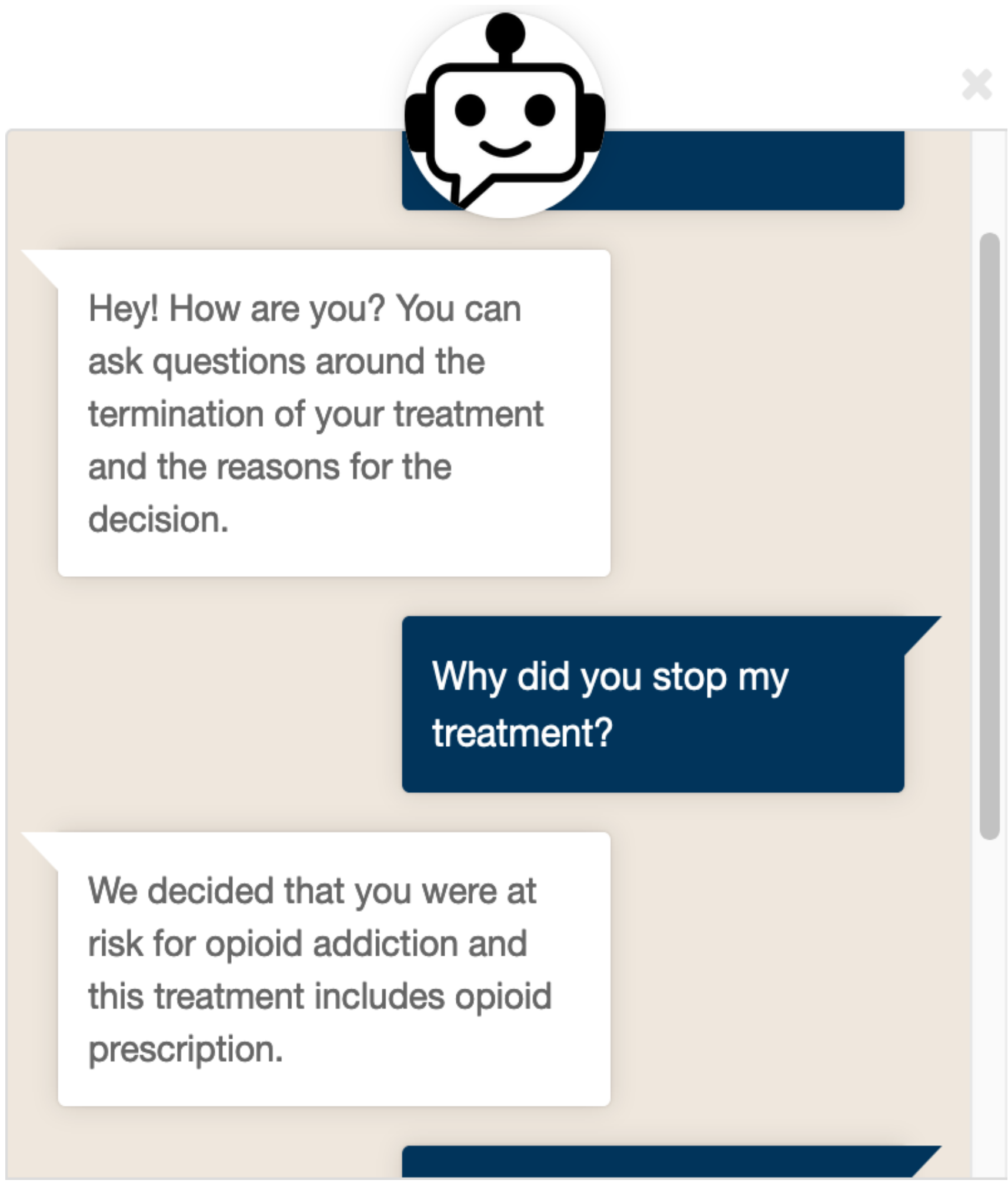
**Empirical Grounding**: Organisations recognize the importance of providing answers, but work is needed to *test and ground their assumptions* about what answers are needed, as their current assumptions are drawn from past experience without AS.

## III. Enabling Answerability in AS by Design

**Mediator agent to connect users with answerable organizations and agents:**

- Quicker response times
- Easier to identify fail points
- Consistency in answers

**Three desired parts of the system:**
1. **Dialogue Representation**
2. **Ontology for Answerability**
3. **Prototype**



Hey! How are you? You can ask questions around the termination of your treatment and the reasons for the decision.

Why did you stop my treatment?

We decided that you were at risk for opioid addiction and this treatment includes opioid prescription.