DeepMind has both a machine learning safety team focused on near-term risks, and an alignment team working on risks from artificial general intelligence. The alignment team is pursuing many different research agendas.

Their work includes:

- Engaging with recent arguments from the Machine Intelligence Research Institute
- The alignment newsletter and podcast, which were produced by Rohin Shah.
- Research like the Goal Misgeneralization paper.
- Geoffrey Irving's work on debate as an alignment strategy.
- "Discovering Agents", which introduces a causal definition of agents, then introduces an algorithm for finding agents from empirical data.

See Shah's comment for more research that they are doing, including a description of some that is currently unpublished.