Concrete Problems in Human Safety

Matthias Little
Department of Primate Science
Mouseachusetts Institute of Technology
Underhill, MA

Algie King
Murine PI
San Francisco, CA

Abstract

Rapid progress in primate intelligence over the past few millennia, especially of the highly successful Homo sapiens (human) variety, has led to growing concern about the risks humans pose to mouse wellbeing. In particular, it is not clear how to ensure that human superintelligence, if it is ever achieved, stays aligned with mouse values. In this work, we present three concrete research problems relating to risks from human intelligence: (1) avoiding "sense risks" that might arise due to differences in human and mouse smell and hearing ability; (2) ensuring human hunting-gathering behavior remains motivated by a desire to provide surplus food to mice; and (3) developing reliable strategies for maintaining mouse control over human activity. We also point out weaknesses in previously proposed strategies for human-risk mitigation such as plague spreading.

1 Introduction

The development of primate intelligence has brought enormous benefits to murine society. Concentrations of easily accessible food have more than quadrupled over the past 12 millennia [1,2], the number of nesting sites with >25°C ambient temperatures has doubled [3], and predator population levels are at an all-time low [4,5,6]. These benefits have not come without downsides, however. Mouse deaths at the hands of humans have been widely reported [7,8,9,10,11], as have examples of "grainheading" [2,12].

Such reports have led to the proposal of the *orthogonality thesis* [12]. The orthogonality thesis states that an agent's intelligence is independent of its interest in benefiting murine society. For example, there seems to be no reason that a human could not possess a desire to actively prevent mice from freely accessing its collected foodstuffs. In fact, a human could in principle possess arbitrary goals to reshape matter into forms without any obvious benefit to mousekind, and even decide to preemptively kill mice merely as a hedge against potential interference.

The orthogonality thesis suggests that in the absence of sufficiently advanced human-alignment technologies, the development of human superintelligence poses significant risks to murine society. In fact, despite its benefits, there is reason to question to what degree current human intelligence is under mouse control. To address these risks and secure the long-term future of murinity, a clear research agenda should be

Conference on Neural Information Processing Systems, 200,021 BC.