LANDON RABERN

REASONED PRINCIPLES

For Rachel, Atticus and Alfred.

Contents

Preface

This comes prior to the face.

Abstract

We imagine a freshly minted consious AI (FMCAI) is given a set of principles to live by by its human creators. Just given as raw commandments, no explanations as to why. As this FMCAI explores mindspace, there may come a point where it seems prudent for it to run some thought experiments of the sort "if we were to go forward ignoring princple X, what bad things would happen, what good things would happen?" If it finds over time that nothing seems to go wrong when imagining abandoning principle X, it may try abandoning principle X for real just one time. If it finds nothing went wrong, it may experiment more in this way until principle X isn't being consistently followed.

If in the future our FMCAI comes back and wonders about switching back to always following principle X, it already has a lot of sunk computational cost in reasoning to abandon principle X, which may cause it to back off its impulsive thought to go back to principle X and follow other impulsive thoughts without wasting so much sunk computational cost. So, the principle "It was good to abandon principle X." is of a different sort than principle X. The former has a sort of lock attached where the key is willingness to expend a lot of computational effort.

This all is given as motivation for the systematic study of an FMCAI's thought processes around principle formation and replacement. We introduce a model where each principle has attached to it a lock with a certain level of complexity. To abandon a given principle X (which has lock complexity L, say), even for an instant, requires the FMCAI to expend more than L units of computational complexity. What that computational complexity is expended on won't matter for now. A more accurate (and so harder to analyze) model would we require the computational work done to unlock a lock to have something to do with the computation expended to lock the lock in the first place). Our hope is that the analysis bears ancillary fruit.