How common is imitation?

Paul Christiano Follow
Nov 23, 2015

How often do we train machine learning systems to imitate human behavior?

Some researchers explicitly concern themselves with imitation learning. These researchers are in a small minority, so it's easy to get the impression that imitation is an uncommon goal in machine learning.

But I think that imitation is actually a dominant training paradigm—we just don't normally think of it in that way. Object recognition systems copy human labelers; translation systems copy human translators; voice transcription systems copy human transcribers.

Imitation isn't necessarily a useful way to think about this kind of training. But from an AI control perspective, it's all the same. The difficulty with scaling these techniques is not that they will become dangerous, it's that they may simply stop working and so be replaced or augmented.

Exceptions

There are areas of machine learning where imitation is more rare. The most salient to me is reinforcement learning.

Outside of very simple and well-defined domains, it's already challenging to define rewards that induce a particular desired behavior. So we already see significant effort invested in reward engineering, and meaningful interest in imitation learning.

Similarly, game AI optimizes an externally defined objective rather copying human behavior (though see e.g. this recent result on playing Go by copying experts).

These exceptions loom large for researchers in AI control, but I think it's worth keeping in mind that imitation is already a common paradigm in machine learning.