SCIE300 Reading Assignment 7 (RA7)

60 points

Question 1 (15 points)

In chapter 12, Bostrom states that the "wireheading syndrome remains a likely outcome" for typical reinforcement learning agents. Look back to chapter 8, where the concept of "wireheading" is defined, and remind yourself of this definition. Then write a few sentences explaining in more detail how the wireheading concept applies to a reinforcement agent.

Question 2 (10 points)

Describe one or more weaknesses of the "value learning" approach described by Bostrom in chapter 12.

Question 3 (10 points)

In the final paragraph of chapter 15, Bostrom suggests that the control of superintelligence may be "the essential task of our age." Do you agree? Give reasons for your answer.

Question 4 (5 points)

Boden's paper depends crucially on a philosophical concept called *intentionality*. Look up the definition of this concept in the Stanford Encyclopedia of Philosophy. In one or two sentences of your own words, give a definition of intentionality.

Question 5 (10 points)

Consider one of Searle's claims, which we could paraphrase in the following way:

> Just as photosynthesis can only occur using a biological substrate, intentionality can only be produced by a substance with causal powers, such as neuroprotein.

Explain in your own words how Boden rebuts this claim.

Question 6 (10 points)

In the last few pages, Boden gives an argument that computer programs can represent understanding and meaning. (Feel free to skim over pages 100-101, which are rather technical, but the remainder of the argument should be comprehensible.) Explain this argument in a few sentences of your own words.

Question 7 (10 points)

Whose argument do you find more convincing, Searle or Boden? Explain your answer in a few sentences.