Defining AI wireheading

by **Stuart Armstrong** 4 min read 21st Nov 2019 6 comments

8

What does it mean for an AI to wirehead its reward function? We're pretty clear on what it means for a human to wirehead - artificial stimulation of part of the brain rather than genuine experiences - but what does it mean for an AI?

We have a lot of examples of wireheading, especially in informal conversation (and some specific prescriptive examples which I'll show later). So, given those examples, can we define wireheading well - cut reality at its joints? The definition won't be - and can't be - perfectly sharp, but it should allow us to have clear examples of what is and what isn't wireheading, along with some ambiguous intermediate cases.

Intuitive examples

Suppose we have a weather-controlling AI whose task is to increase air pressure; it gets a reward for so doing.

What if the AI directly rewrites its internal reward counter? Clearly wireheading.

What if the AI modifies the input wire for that reward counter? Clearly wireheading.

What if the AI threatens the humans that decide on what to put on that wire? Clearly wireheading.

What if the AI takes control of all the barometers of the world, and sets them to record high pressure? Clearly wireheading.

What if the AI builds small domes around each barometer, and pumps in extra air? Clearly wireheading.

What if the AI fills the atmosphere with CO_2 to increase pressure that way? Clearly wire... actually, that's not so clear at all. This doesn't seem a central example of wireheading. It's a failure of alignment, yes, but it doesn't seem to be wireheading.

Thus not every example of edge or perverse instantiation is an example of wireheading.

Prescriptivist wireheading, and other definitions

A lot of posts and papers (including some of mine°) take a prescriptivist approach to wireheading.

They set up a specific situation (often with a causal diagram), and define a particular violation of some causal assumptions as wireheading (eg "if the agent changes the measured value X without changing the value of α, which is being measured, that's wireheading").

And that is correct, as far as it goes. But it doesn't cover all the possible examples of wireheading.

Conversely, this post°defines wireheading as a divergence between a true utility and a substitute utility (calculated with respect to a model of reality).

This is too general, almost as general as saying that every Goodhart curse is an example of wireheading.

Note, though, that the converse is true: every example of wireheading *is* a Goodhart curse. That's because every example of wireheading is maximising a proxy, rather than the intended objective.

The definition

The most intuitive example of wireheading is that there is some property of the world that we want to optimise, and that there is some measuring system that estimates that property. If the AI doesn't optimise the property, but instead takes control of the measuring system, that's wireheading (bonus points if the measurements the AI manipulates go down an actual wire).

This re-emphasises that "wireheading is in the eye of the beholder°": if our true goal is actually the measuring system (maybe our AI is in competition with another one to maximise a score in a game, and we really don't care how it does this), then there will be no wireheading, just an AI following a correct objective.

Thus wireheading is always a failure of some (implicit or explicit) goal; thus every example of wireheading is a failure of value alignment, though the converse is not true.

Also key to the definition is the fact that the measuring system is, in some sense "much smaller" than whatever property of the system it is measuring. Pumping out CO_2 is not the correct instantiation of some goal along the lines of ″increase air pressure so humans enjoy better weather″; but nor is it merely manipulating the measurement of that goal.

The definition

Thus we can define wireheading as:

- Given some implicit goal G, an agent wireheads if, instead of moving towards G, it manipulates some *narrow* measurement channel that is intended to measure G, but will fail to do so after the agent′s manipulation.

The difference with the prescriptivist approach is that the measurement channel is not specified; instead, we ask whether we can usefully characterise some feature of the setup as a ″narrow measurement channel″, and then apply the definition.

This can be seen as a particular failure of abstraction: the abstract goal G was collapsed to the output of the measurement channel.

Examples, counter-examples, and edge cases

Under this definition, all the intuitive examples of wireheading above fit: the measurement channel the AI takes over (its internal counter, the wire going into it, the statements made by humans, the barometers, the immediate environments of the barometers) is always much smaller than the whole atmosphere, which was its intended goal.

And that′s why the CO_2 example isn′t wireheading: the AI is doing a massive manipulation of the world, on the same scale as its intended goal; it isn′t just manipulating the measurement channel[1].

The case of the domes around the barometers is an interesting one to consider. Obviously, if the AI put a dome around the planet and pumped in extra air, this wouldn′t count as wireheading. Thus, we can imagine the domes growing bigger and merging, thus giving a smooth transition from ″clearly wireheading″ to ″clearly not wireheading″, and showing that ambiguous cases must exist.

We can also produce examples of Goodhart curse that are not wireheading. Take the practice of ″teaching to the test″. In this case, there is a simple objective (the test results) and the school acts to optimise for that objective. However, in typical schools this is not wireheading;

teaching to the test involves drilling students in specific skills, training them, and having them memorise certain facts. Though these are done specifically to pass the test, these are the kinds of actions that a teacher would undertake anyway. One can talk about how this ″narrows″ the intellect, but, except in extreme cases, this cannot be characterised as gaining control of a narrow measurement channel.

For an interesting edge case, consider the RL agent playing the game CoastRunners. As described here, the score-maximising agent misbehaved in an interesting way: instead of rushing to complete the level with the highest score possible, the agent instead found a way to boat in circles, constantly hitting the same targets and ever increasing its score.

Is that wireheading? Well, it′s certainly Goodhart: there is a discrepancy between the implicit goals (got round the course fast, hitting targets) and the explicit (maximise the score). But do we feel that the agent has control of a ″narrow″ measurement channel?

I′d argue that it′s probably not the case for CoastRunners. The ″world″ for this agent is not a particularly rich one; going round and round and hitting targets is what the agent is intended to do; it has just found an unusual way of doing so.

If, instead, this behaviour happened in some subset of a much richer game (say, SimCity), then we might see it more naturally as wireheading. The score there is intended to measure a wider variety of actions (building and developing a virtual city while balancing tax revenues, population, amenities, and other aspects of the city), so ″getting a high score while going round in circles″ is much closer to ″controlling a measurement channel that is narrow (as compared to the implicit goal)″ than in the CoastRunners situation.

But, this last example can illustrate the degree of judgement and ambiguity that can exist when identifying wireheading in some situations.

[1] Note that the CO_2 example can fit with the definition of this post ◦. One just needs to imagine that the agent′s model does not specify the gaseous content of the air in sufficient detail to exclude a CO_2-rich air as a solution to the goal.

This illustrates that the definition used in that post doesn′t fully capture wireheading. ↩