Robot Revelations

By Jeff Huggins

Imagine yourself designing and building a robot.

You've *almost* completed your creation: You only need to install the final three control circuits into your robot's operating system—i.e., its *brain*—and program them into its hierarchy of priorities.

Oh, and: Your robot, of course, uses energy in order to work. It's a rechargeable electric robot. When its battery runs low, and has only an hour's worth of electricity left, the battery sends a signal to the control center of the robot's operating system: "running low!".

Now, you have three simple control circuits left to install:

When one control circuit gets the "go" signal, it tells the robot to mow your lawn.

When another control circuit gets the "go", it tells the robot to play chess.

When a third control circuit gets the "go" signal, it tells the robot to plug its built-in battery into an electrical outlet to get recharged.

Your final task—should you choose to accept it and complete your creation—is to install these three circuits and program your robot's hierarchy of priorities.

One of several things you'll need to decide is this: When your robot's battery sends the "running low!" signal to its operating system, what should your robot do first, with its remaining power: mow the lawn, play chess, or find an outlet and plug itself in?

Hhhhmmmm.

First, consider the question from your own standpoint, i.e., from the standpoint of the robot's designer. If you want your robot to "live", of course, you'll program it to find an outlet, plug itself in, and recharge, and then perhaps to mow the lawn and play chess later. On the other hand, if you want your robot to "die", you'll program it to mow your lawn or play chess when it receives the "running low!" signal. Or, if you don't really care one way or another, you might flip a coin to decide, or sell your robot as scrap metal.

Now imagine: What if the robot is to (somehow) design itself from scratch? In other words, what if there is no designer or if the designer is out of the picture for these final steps? If you can playfully imagine, for the moment, that the robot can design itself, would it be more reasonable and sensible for the robot to program itself to recharge, mow the lawn, or play chess, when it gets the "running low!" signal?

Of course, robots probably can't design themselves from scratch, so let's move one more step into the exercise: What if the robot *inherited* itself? Put another way, what if the robot was born equipped with an operating system that roughly tells it, usually, to recharge itself first—before mowing lawns or playing chess—whenever it gets the "running low!" signal?

Now, in this context, consider these questions for a moment:

- Which of the following would or should a sensible robot do?
    - Respect and affirm its existing operating hierarchy with respect to placing a priority on recharging itself before mowing lawns or playing chess, upon receiving the "running low!" signal.
    - Rewire itself to mow lawns or play chess upon receiving the "running low!" signal.
    - Flip a coin to decide.
- In your view, and remembering again that you are approaching the matter from the standpoint of the robot now, which of the following ingredients and considerations *informed* the *choice* you made above?
    - Solely emotion and self-love—because excellent reasoning cannot help one choose between the three options facing me, as the robot.
    - Solely excellent reasoning—in some sort of "pure" form, i.e., disembodied from emotion.
    - An intimate combination of emotion *and* excellent reasoning, with the "excellent reasoning" component playing a substantial role in being able to differentiate between the three options, weigh them relative to each other, and choose one that, under the circumstances, is more consistent with key considerations of "reason" than the other two.

When considering these matters, it also helps to recognize several additional factors that might or might not be obvious at first:

First, if you, in your initial role as an autonomous designer, did *not* want your robot to live, you would (presumably) have not programmed it to recharge upon receiving the "running low!" signal. You might have programmed it to ignore the signal, or to mow lawns, or to play chess. Or, why build a robot at all? Of course, these conclusions are not *necessarily* so. But, if a world of robots that *do* know to recharge before playing chess somehow came about, that would be a pretty good sign, of course, of one of three things: either that their designer designed them *to* recharge themselves; or that their designer did (or does) not care one way or the other, but at least *allowed* them to develop and inherit a *propensity* to recharge; or that there is no designer.

Second, aside from the whole issue of a designer, and coming at the matter from the standpoint of the robot itself, notice that the robot would have to have active reasons *to* rewire itself (to play chess or mow lawns rather than to recharge) that were *more compelling than* those to leave its wiring *as is* (with respect to this particular priority) in order to have "reasoned justification" to rewire itself. Put another way, the robot's *status quo* programming is the *default* unless and until compelling reasoning could cause the robot to *reject* its inherited priority to recharge before mowing or chess.

Third, "justification" and "reasoning" are, of course, two different things (in many senses anyhow), and the relationships and differences between the two depend on what a person means by the terms and/or on the question being addressed. For example, we shouldn't assume that an absolute authority independent of ourselves (humans) will "justify" our existence *to* us and *for* us, tell us what *choices* to make, or tuck us in at night. But, that *doesn't* mean, of course, that the very best of human reasoning can't weigh three options and inform us of which one of the three is *most* consistent with the considerations of "reason".

(Of course, the robot could also be programmed to generate little copies of itself before it gets too rusty, stuck in the mud, or squashed by a big boulder. In the interests of simplicity and of maintaining a "G" rating, I've limited the present illustration to the merely electrical.)

Finally, three more questions to consider:

- What does this exploration suggest about the matters of "is" and "ought"?

- What does it suggest about "reason", "justification", and related matters?
- What does it suggest about the differences and relationships between knowledge and wisdom?

Now where's that outlet?

