The Problem of Original Agency

Don Berkich

Texas A&M University-Corpus Christi

Abstract

The problem of original intentionality–wherein computational states have at most derived intentionality, but intelligence presupposes original intentionality–has been disputed at some length in the philosophical literature by Searle, Dennett, Dretske, Block, and many others. Largely absent from these discussions is the problem of original agency: Robots and the computational states upon which they depend have at most derived agency. That is, a robot's agency is wholly inherited from its designer's original agency. Yet intelligence presupposes original agency at least as much as it does original intentionality. In this talk I set out the problem of original agency, distinguish it from the problem of original intentionality, and argue that the problem of original agency places as much of a limit on computational models of cognition and is thus at least as vexing as the problem of original intentionality.

Introduction

In this talk I argue that there is to the problem of original intentionality a parallel yet distinct limitation on computational models of cognition which I shall call the problem of original agency. I begin by describing a thought experiment by Susan Wolf which suggests the problem but ultimately reduces to the problem of original intentionality. Having briefly described the problem of original intentionality, I then construct a further thought experiment, the Case of the Paranoid Roboticist, to isolate and describe the problem of original agency. I close by explaining why I think this is a serious problem for philosophers and computer scientists alike quite independently of the problem of original intentionality.

Wolf's Perfect Android

In *The Importance of Free Will*, Susan Wolf (1993) invites us to compare what she calls 'reactive' attitudes with the 'objective' attitude. Reactive attitudes are just those attitudes we take towards one-another – the feelings, in particular, of gratitude or resentment we have in response to others' actions. The objective attitude is the attitude a mechanic might take with respect to a car. The behavior of the car is not something it makes sense to appreciate or resent. A car that won't start might anger us, but the thought of punishing the car is absurd.

Wolf's broader project is to determine whether or not it makes sense – whether or not it would be rational – to take reactive attitudes towards a putative morally responsible agent even after it has been found out that the agent is psychologically and/or physically determined to do precisely he or she did. In the course of her discussion, Wolf presents the case of the perfect android.

To all appearances, behavioral and otherwise, Wolf's perfect android is a human being. Despite appearances, it is an android: It is completely programmed in the sense that its actions are

programmed, its choices are programmed, and even its cognitive processes, if it can be said to have them, are entirely programmed. Nothing it does or 'thinks' is unexpected by its programmer. Indeed, to ensure that this is so the programmer is actively and continually attached to the android so as to program its responses on “a day-to-day or moment-to-moment basis.” (p. 110) A puppet is perhaps the clearest analogy to what Wolf has in mind, although as Wolf puts it, “[o]ne might imagine the relation between robot and programmer to be very much like a possible relation between author and character; or, perhaps even better, one might imagine the relation to be like the relation between a magician and a human being over whose thoughts and bodily movements the magician has complete control.” (p. 110)

Wolf concludes that taking the objective attitude with respect to the android is the only way to remain consistent with the facts. In no sense relevant to justifying reactive attitudes can her android be considered a responsible agent.

> In light of the nature of the robot's programming, I believe that the only way of living in accordance with the facts would be by regarding the robot solely with the objective attitude. That is, I believe that the robot is not a free and responsible being in whatever sense of 'free and responsible' the objects of our reactive attitudes are ordinarily assumed to be. Were we to be purely rational, we would allow ourselves to feel some emotions toward the robot, but we would not feel those emotions or sentiments constitutive of our reactive attitudes. For though the robot might choose to perform the actions he performs, he chooses to perform them only because he is programmed to so choose. Though his decisions and judgments may be preceded by thoughts which look or sound like reasons, he cannot be said to reason to these conclusions in the way we do. He is not in

> ultimate control of his value, his personality, or his actions. He is, properly speaking, only a vehicle for carrying out the plans (if plans there be) of his programmer. (p. 110)

I find Wolf's perfect android puzzling. The android is physically and behaviorally indistinguishable from an ordinary human person. Wolf's claim is that her android is programmed, but her conception of programming stretches what we ordinarily understand programming to be. The program is not written first and then executed by the android. Wolf's programmer must issue instructions on a moment-to-moment basis. In order for this to work Wolf's programmer would have to have moment-to-moment access to what the android 'sees', 'feels', 'hears', etc. A feedback-instruction loop would have to exist between the android and the programmer for the android to have any chance whatsoever of passing as an ordinary human person. Given this picture, it is hard to understand why one should accept the conclusion that "the only way of living in accordance with the facts would be by regarding the robot solely with the objective attitude."

To make sense of Wolf's story, it would have to be the case that the android is essentially a humanoid ROV (Remotely Operated Vehicle). Current ROV's employ a stereoscopic camera system which links to the operator's stereoscopic goggles so as to give the operator the depth perception and visual experience of actually being on the ocean floor or in the nuclear reactor. Wolf's 'programmer' is not so much a programmer as an operator. As such, it is perfectly appropriate to take reactive attitudes with respect to the android, since the android's agency just is its operator's agency. One's resentment of or gratitude towards the android just is one's resentment of or gratitude towards the android's operator.

Of course, if the operator murdered via the android, it would make little sense to imprison the

android and leave the operator at large unless the operator's sole mode of agency were the android. If the operator were a brain in a vat wired to radio controllers in such a way that it sees via the android's cameras, hears via the android's microphones, feels via the android's tactile transducers, and acts via the android's motors, then imprisoning the vatted brain while leaving the android at large would make as little sense as imprisoning the android while leaving the operator free would make in the former case. Yet this only serves to underscore the point that Wolf's android wholly inherits its operator's agency and highlight the oddity of Wolf's claim that it would only make sense to take the objective attitude towards her android.

What if the android were programmed – in the usual sense of the word – instead of being operated? The programmer codes a set of instructions, stores them in the android, and has nothing more to do with the android; the android executes the instructions without any further meddling by the programmer. Wolf's assumption that the android would be behaviorally indistinguishable from ordinary human persons is, in this case, egregiously question-begging, since it implies that the android would be quite as capable as any human person at passing the Turing Test for artificial intelligence. Thus there would be no reason whatsoever to adopt the objective view with respect to the android, since whatever makes us think that we are responsible agents is equally true of the android. On the other hand, if the android were behaviorally distinguishable from a human person, then Wolf's conclusion that it is rational to adopt the objective attitude with respect to the android is plausible. What also follows is an argument against machine agency.

1. If X is an agent, then it is not rational to take the objective attitude with respect to X.
2. It is rational to take the objective attitude with respect to machines.

Hence, 3. Machines are not agents. 1&2

Notice that this argument does not conclude that machines cannot be agents, only that machines are not agents since, to date, it is rational to take the objective attitude towards them. Adding the appropriate modal modifiers raises an interesting question.

1. Necessarily, if X is an agent, then it is not rational to take the objective attitude with respect to X.
2. Necessarily, it is rational to take the objective attitude with respect to machines.

Hence, 3. Machines cannot be agents. 1&2

Why should we think that, *necessarily*, it is rational to take the objective attitude towards machines? A good reason might be if machines could never pass the Turing Test: Any argument against the possibility of machine intelligence is, *a fortiori* and unsurprisingly, an argument against the possibility of machine agency. Wolf's perfect android does not therefore introduce any *special* problem for computational models of cognition apart from the usual problems associated with machine intelligence – viz. the problem of original intentionality, in particular.

Original Intentionality

The screen upon which I am now focused has, in addition to this sentence, a series of system status indicators off to the side which give me lots of useful information. They tell me, for instance, that the motherboard is currently at 84 degrees and the processor cooling fans are turning at a quiet 5000 r.p.m. They indicate which processes are currently running and the extent

to which the twin CPUs are being taxed by those processes. They even give me current weather information from the local airport (temperature 71°F, pressure 29.97, humidity 61%, and winds out of the southeast at 5 mph), the date and time, and which mp3 is currently playing.

One might be inclined to conclude that my computer is in one respect just like me: we are both intentional systems. That is, the computer has indicators and states which are about other things – the weather, the time, sentences on the problem of original intentionality, etc – just as I have mental states such as my beliefs about the weather, the time, and certain arguments. Yet very much unlike me, it seems the computer has at most derived intentionality. As Haugeland (1997) puts it,

> Here's the idea: sentence inscriptions–ink marks on a page, say–are only "about" anything because we (or other intelligent users) mean them that way. Their intentionality is second-hand, borrowed or derived from the intentionality that those users already have. ...Our intentionality itself, on the other hand, cannot be likewise derivative: it must be original. ('Original', here, just means not derivative, not borrowed from somewhere else. If there is any intentionality at all, at least some of it must be original; it can't all be derivative.) (p. 7)

Why is this a problem for AI? Haugeland explains:

> The problem for mind design is that artificial intelligence systems, like sentences and pictures, are also artifacts. So it can seem that their intentionality too must always be derivative–borrowed from their designers or users, presumably–and never original. Yet, if the project of designing and building a system with a mind of its own is ever really to succeed, then it must be possible for an artificial

system to have genuine original intentionality, just as we do. (1997, p. 7)

Philosophers have spent a great deal of ink wrestling with the problem of original intentionality and even wrestling with the problem of whether it is a problem in the first place. There is, however, a parallel and, I shall argue, even more vexing problem.

Original Agency

Wolf's perfect android presents no special problems apart from familiar ones like the problem of original intentionality, but it does suggest a special problem. It makes sense to take the reactive attitude towards Wolf's android, I argued, since its agency just is, in toto, its operator's agency – this much is obvious when we appreciate that Wolf's perfect android is nothing more than a humanoid ROV. Contrary to Wolf, anger at the perfect android's pranks is perfectly justified since its pranks just are the operator's; yelling at the perfect android just is to yell at the operator. An excellent question at this point is whether something similar would be true of a programmed, as opposed to operated, android. Does a programmed android also inherit its programmer's agency? In what follows I argue that it does, and necessarily so.

Suppose that Ted the Survivalist, in a fit of deepening paranoia, resolves to keep his gun trained on the door of his shack so as to kill anyone who might try to enter. After twenty hours of this Ted frightens himself by startling awake upon nodding off: for a few seconds at least he was vulnerable! As clever as he is paranoid, Ted fashions a simple system of strings and pulleys such that the gun fires dead-center into the doorway when the door is opened. Ted is free to sleep and go about his usual business, secure in the knowledge that anyone trying to enter his shack will be killed.

Ted's contraption acts so as to kill any would-be attackers. It has this potential only insofar as Ted built it thus and so. The agency of the contraption, should it ever fire, is thus wholly derived from Ted's original, albeit deranged, agency. In that sense it would be far better to put Ted in a secure psychiatric hospital than the contraption if, say, the postman were killed.

Let us press the example further. Suppose that Ted isn't any ordinary paranoid survivalist. Ted is also a brilliant roboticist with considerable economic resources. To protect himself, Ted constructs a mobile adaptive live-fire robot, anticipating somewhat recent DARPA advances. The robot, which Ted affectionately dubs 'R2D3', looks like a wheeled trash-can. It has three arms stuck out the sides that articulate in four locations and terminate in large, fully-automatic guns. On top of R2D3 stands a thin, retractable pole, and on the top of the pole is the its 'head'. The 'head' is just a pair of side-by-side cameras which can swivel in nearly every direction. Ringing R2D3's base are sonar sensors which allow it to navigate from room to room and around the yard.

In operation, R2D3's head continuously bobs up and down and swivels back and forth as it scans its vicinity. R2D3's head orients on any movement and, using cues such as bi-lateral symmetry, zeroes in on any faces. It then compares key features of the face to an on-board database of such features. If there is, within a certain narrow tolerance which Ted keeps notching up as his paranoia deepens, a match in the database, then the object the robot is tracking is a “friend”. If it fails to find a match, the object is a “foe”, at least for a few milliseconds.

Fortunately R2D3 is adaptive in the sense that it updates its “friends” database whenever Ted himself opens the door and speaks in normal tones with a person. Thus Ted congratulates himself on having protected the postman – until, that is, the postman falls ill and his substitute walks into the yard.

Suppose R2D3, suffering no malfunction whatsoever, kills the substitute postman. Is its killing of the substitute postman an example of derived or original agency? Put another way, does it make any more sense to put R2D3 in the secure psychiatric hospital than Ted's original string-and-gun contraption? Surely not. Of course R2D3 ought to be disabled, but not as a punitive measure. We disable the robot for precisely the same reason that we disable the string contraption: to avoid any more 'accidents'. The scrap heap is the appropriate end for R2D3. Ted, the lethal robot's designer and programmer, is the one who gets to go to the secure psychiatric hospital.

Machines are designed and programmed; their agency can never be original, yet agents require original agency. Setting the argument out, we have:

	1.	Necessarily, if X is an agent then X has original agency.	
	2.	Necessarily, if X is designed and programmed, then X has only derived agency.	
	3.	Necessarily, if X has only derived agency then X does not have original agency.	
	4.	Every machine must be designed and programmed.	
Hence,	5.	No machine can be an agent.	1,2,3&4

It may be objected that some machines are designed and programmed by other machines. This is true, but then all that can be said is that any such machine will only have derived agency two or more times removed. It might also be objected that some machines are not programmed at all: the algorithms by which they operate are developed by reward and punishment or even, in the case of genetic algorithms, by directed 'evolution'. Such machines still have only derivative agency – derived, in these cases, from the rewarder, the punisher, or the one who directs the

evolution.

Finally, it may be objected that unpredicted behavior, even from the standpoint of the machine's designer, or surprisingly reasonable behavior as would have been the case were R2D3 to have refrained from killing the substitute postman warrant the attribution of original agency. Complicated and adaptive design is nonetheless designed. The mere fact that the designer herself is surprised only accrues praise for the designer, never her machine. Nor is the problem of original agency without consequences for philosophy: If machines can have at most derived agency, then discussions of robot ethics, for example, are mostly otiose.

References

Haugeland, John. 1997. What is Mind Design? In J. Haugeland (ed.) *Mind Design II: Philosophy, Psychology, Artificial Intelligence*. Cambridge, Mass.: MIT Press. A Bradford Book pp. 1-28.

Wolf, Susan. 1993. The Importance of Free Will. In J.M. Fischer and M. Ravizza (eds.) *Perspectives on Moral Responsibility*. New York: Cornell University Press pp. 101-118.