Autonomous Robots*

Don Berkich

5/29/08

Texas A&M University-Corpus Christi

Abstract

"Autonomy" enjoys much wider application in robotics than philosophy. Roboticists dub virtually any robot not directly controlled by a human agent "autonomous" regardless of the extent of its behavioral repertoire or the complexity of its control mechanisms. Yet it can be argued that there is an important difference between autonomy as not-directly-controlled and autonomy as self-control.

With all due respect to the enormous achievement Mars rovers Spirit and Opportunity represent, in this paper I argue that the roboticists' conception of autonomy is over-broad in the sense that it is too easily achieved, uninteresting, and ultimately not especially useful. I then propose a theory of autonomous agency that capitalizes on the roboticists' insights while capturing the philosophical conception of autonomy as self-rule. I close by describing some of the capacities autonomous agency so construed presupposes and explaining why those capacities make it a serious challenge to achieve while arguing that it is nonetheless an important goal.

word count: 3318

Introduction

In a 1986 manuscript Rod Brooks set the path for robotics development by announcing that he wished "to build completely autonomous mobile agents that co-exist in the world with humans, and are seen by those humans as intelligent beings in their own right." (1999 p. 86) Brooks astutely sided-stepped difficult questions about intelligence and intentionality with the decidedly dennettian phrase, "are seen by", yet showed not the slightest hesitation in declaring the goal to be autonomous agency. (Dennett 1987) To be sure, the field of situated robotics Brooks arguably initiated has achieved far more than the commercial success of the floor cleaning Roomba: Mars rovers Spirit and Opportunity are particularly spectacular examples. Nor should Brooks, writing what amounts to more of an advertisement than a closely argued essay which explicitly shrugs off philosophy--he admits to having "no particular interest in the philosophical implications..., although clearly there will be significant implications" (1999 p. 86)--be held to the exacting standards of philosophical analysis.

Philosophical implications notwithstanding, perhaps Brooks ought to be interested in his philosophical *assumptions*. His sweeping assumption that autonomous machine agency is possible strikes one as desperately needing justification in light of the intuitive rub, if not outright conflict, between *mechanism* on the one hand and *autonomy* on the other. It is one thing to suppose that a machine can act – a questionable assumption in its own right – quite another to assume that a machine, the paradigm of causal pawns, can act of its own accord.

In this paper I examine the capacities Brooks' mobile agents would have to enjoy to be completely autonomous in any interesting sense of the phrase. I begin by contrasting the use of "completely autonomous" in computer science circles with its philosophical use and argue that its use in computer science is by comparison over-broad and uninteresting. I then explain how one

might go about analyzing "completely autonomous" in such a way as to capture its philosophical use. I leave as an open question whether *any* agent, human or robot, could be autonomous in this sense, but in closing I explain why roboticists should care.

Two Conceptions of Autonomy

Philosophers have no monopoly on the term "autonomous", yet computer scientists--who more frequently come to computer science from mathematics than philosophy--appear to have a much broader conception of autonomy. To help clarify his goal of achieving completely autonomous mobile agents, Brooks goes on to explain that his robots, which he calls 'creatures', would be adaptive, robust, opportunistic, and active. Thus quoting,

- A Creature must cope appropriately and in a timely fashion with changes in its dynamic environment.
- A Creature should be robust with respect to its environment; minor changes in the properties of the world should not lead to total collapse of the Creature's behavior; rather one should expect only a gradual change in capabilities of the Creature as the environment changes more and more.
- A Creature should be able to maintain multiple goals and, depending on the circumstances it finds itself in, change which particular goals it is actively pursuing; thus it can both adapt to surroundings and capitalize on fortuitous circumstances.
- A Creature should do something in the world; it should have some purpose in being. (1999 p. 86)

Two points bear emphasis. First, situated roboticists seek to create AI on a par with insects by constructing mobile robots with sensors for immediate feedback about the results of their motions. The idea is to construct robots which behave in such a way that they are able to negotiate their environment; the goal in situated robotics is not the intelligence of the situated robot *per se* but on its ability to succeed in its environment. The successful situated robot adapts to its environment by 'learning' strategies for accomplishing tasks. It thereby uses the environment, not its program, as its source of information about the environment. Presumably the more adaptive

and successful robot will also be the more intelligent robot.

Second, the situated robot is not pre-programmed with detailed instructions for completing its tasks. Unlike industrial roboticists, who spend nearly as much time specifying the environment in which their robots will operate as designing the robots themselves, the situated roboticist in all likelihood has no way of knowing in advance what the situated robot will encounter and do in a given environment apart, perhaps, from global directives like "move from position P to position P*." The roboticist has no way of knowing exactly how the situated robot will end up solving the problem of a rock placed in its path, for instance. An excellent example of this was the successful use of situated robotics techniques in the Mars mini-rover, Sojourner Rover (SR). Because of a substantial delay in communication time between earth control and SR, global objectives like "move to rock C" could be given to SR while the problem of just how to get to C around intervening rocks A and B was left up to SR itself. Mission control could not always anticipate the path SR would end up taking.

A robot, in short, is completely autonomous provided it is not under the *direct* control of a human operator. Thus a bomb-disposal Remotely Operated Vehicle is not autonomous inasmuch as it is tethered by joystick to a human operator while the SR and the Roomba presumably are autonomous because they do what they do without direct human intervention. Call autonomy of the not-directly-controlled variety “autonomy-as-independence”. Perhaps unsurprisingly, autonomy-as-independence is consistent with the mathematical notion of autonomy, whereby a differential equation is autonomous if the behavior of the system it describes is translation invariant over time and in that sense independent.

Autonomy-as-independence contrasts sharply with the more etymologically consistent philosophical conception of autonomy-as-self-rule. As Dworkin puts it,

> What I believe is the central idea that underlies the concept of autonomy is indicated by the etymology of the term: *autos* (self) and *nomos* (rule or law). The term was first applied to the Greek city state. A city had *autonomia* when its citizens made their own laws, as opposed to being under the control of some conquering power. There is then a natural extension to persons as being autonomous when their decisions and actions are their own; when they are self-determining. (1988 pp. 12-13)

To be sure, "self-determining" needs unpacking and, since our interest is in completely autonomous *robots*, the unpacking must be computationally tractable. Suffice it to say at this point that many more things are autonomous in the autonomy-as-independence sense than the autonomy-as-self-rule sense. Bacteria, bump-and-turn toy cars, and Roomba's are equally independent of the direct control of some other agent. A suitably equipped bacterium is capable of independent motion, yet it is completely at the mercy of the stimulus and response mechanisms for which it has been selected by evolution. It cannot but follow a chemical gradient, for example. Similarly, the Roomba can navigate and optimize floor-cleaning patterns, but in doing so it merely follows a somewhat more general set of rules than those implemented in the construction of the bump-and-turn toy car.

The point is that a mobile agent can be independent of any direct control yet be at the complete mercy of internal control mechanisms. Put simply, it is not enough that a robot be able to react in an appropriate way to a stimulus, since the reaction and, presumably, its appropriateness may simply be an artifact of its design and programming (even with the addition of learning algorithms or bayesian decision algorithms.) A narrower and, it seems to me, much more interesting sense of autonomy is one in which the agent's internal contrd mechanisms are themselves subject to the agent's control, if only because of the vastly enhanced adaptive control such an agent presumably exercises.

Autonomy-as-Self-Rule

It may be, of course, that complete autonomy-as-self-rule is simply not achievable by any organism. Perhaps human agents at best approximate autonomy-as-self-rule. Precisely what capacities would an agent have to have to be completely autonomous in the self-rule sense?

Consider first how actions might be distinguished from mere occurrences. The simplest approach is a recursive specification of actions, as follows:

> Let S be an organism, let 'u', 'v', and 'w' be replaced uniformly by verbs, and let t, t*, and t^ be (not necessarily distinct) times.
>
> The Base Axiom
>
> Ax1: S's deliberately u-ing at t is an action.
>
> The Induction Axiom
>
> Ax2: If S's u-ing at t is an action, then S's v-ing at t* is an action if either
>
> S v's at t* by u-ing at t
>
> or
>
> S u's at t by v-ing at t*.
>
> The Closure Axiom
>
> Ax3: Nothing else is an action.

So what distinguishes an action from a mere occurrence, on this account, is its having been sourced in deliberation. Although deliberation begs many questions, an agent is distinct from a mere causal mechanism in having the capacity to shape its contributions to its causal environment. In giving an account of autonomy-as-self-rule which bears on *how* the agent shapes its causal contributions, it will be useful to further stipulate

> The Transitivity Axiom
>
> Ax4: If S u-ed at t by v-ing at t* and v-ed at t* by w-ing at t^, then S u-ed at t by w-ing at t^.

The By-History Definition

Df1: A *by-history* of S's action a_n $=_{df}$ any ordered n-tuple $<a_1,..., a_n>$ of S's actions $a_1,..., a_n$ such that, for each a_j, a_k, S a_k's by a_j-ing.

The Pathway Definition

Df2: A *pathway* $=_{df}$ any segment of a by-history of an action.

Autonomy-as-self-rule presupposes that the agent exercise control over the ways in which it shapes its causal contributions. Thus,

The Axiom of Autonomy

Ax5: S autonomously v's at t* by u-ing at t iff

a) S deliberately desires v-ing at t*,

b) S deliberately believes that v-ing at t* is probable to some non-zero degree d of u-ing at t via a pathway p,

and

c) if not (a) or not (b), then S would not have u'ed at t.

On this view,

> [A]utonomy is conceived of as a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or change these in light of higher-order preferences and values. By exercising such a capacity, persons define their nature, give meaning and coherence to their lives, and take responsibility for the kind of person they are. (Dworkin, 1988, p. 20)

Complete autonomy will not, however, have been established until the agent has the capacity to shape the deliberative basis by which it exercises control over its own deliberations. That is, an agent cannot be said to be *completely* autonomous if it is at the mercy of immutable higher-order preferences and values (and beliefs). If our goal is complete autonomy-as-self-rule, then this is at best *first-order autonomy*. Echoing Frankfurt (1971), an agent enjoys first-order autonomy when it has deliberative control over the deliberative basis it employs in determining its course of action.

Moving beyond Jackson, we may say that the second-order autonomous agent has in addition deliberative control over the deliberations it brings to bear on the deliberative basis it employs in determining its course of action, and so on.

To be sure, deliberation bears a great deal of weight in this account. Without becoming lost in particular deliberative strategies, all deliberations share in common the capacity to contrast alternatives. The capacity to contrast alternatives in turn presupposes the capacity to represent counterfactual states of affairs, and it is this capacity, I submit, which ultimately secures autonomous agency. That is, my raising my arm is not my arm's raising since it, my raising my arm, has a causal history involving my intention to raise my arm. I *deliberately* raised my arm, in short. Deliberately raising my arm, however, presupposes a capacity to form arm-raising relevant beliefs and desires. What distinguishes beliefs from desires, however, is not just the difference in attitude they represent but their admissible representational content. To wit, no desire is ever directed at an actual state of affairs. Whereas beliefs can be about actual and possible states of affairs, desires are exclusively about counterfactual states of affairs. Counterfactual representations can also, however, include as part of their content the very beliefs and desires which rationalize the agent's actions. So just as an agent has the capacity to deliberately raise her arm or not, the autonomous agent has the capacity to deliberately form, or not, arm-raising relevant beliefs and desires. The agent's capacity to reflectively or recursively deliberate upon any of the agent's own propositional attitudes provides at least one way to understand complete autonomous agency in the interesting sense of autonomy-as-self-control.

In sum, a mobile agent is completely autonomous if its actions are n-order autonomous, where the n-th order exhausts all the agents action relevant beliefs and desires. The autonomous agent is something like a deliberative Neurath's Boat, constantly evaluating and revising beliefs

and desires in light of other deliberatively evaluated beliefs and desires.

The capacity to form counterfactual representations of possible states-of-affairs which can include the agent's own states-of-affairs, together with the capacity to form belief, desire, and perhaps other attitudes towards representations so-formed are minimally necessary for autonomous agency on this view. Setting aside the (admittedly huge) problem of attitude formation and robot attitudes for another time, counterfactual representation presents its own special problems. At the very least we need a computationally tractable account of counterfactual representation. One possibility (cf. Cummins 1996; Swoyer 1991) is to account for counterfactual representation in terms of embedded isomorphic relational structures, but using the theory for counterfactual representation raises several crucial questions:

- Complexity: How do we optimize embeddings so as to minimize computational overhead?
- Formability: If possible states of affairs are represented by transformations on embedded isomorphisms, then precisely how and by what factors are those transformations constrained for relevancy?
- Attainability: What metric can be used to assess the action-relevant, agent-relative realizability of counterfactual states of affairs represented by embedded isomorphisms-i.e., in Lewisian terms (Lewis 1973) how does the agent determine the closeness of possible worlds so represented?

In light of these puzzles, perhaps no agent is ever completely autonomous, but only approximates complete autonomy.

The Value of (Complete) Autonomy-as-Self-Rule

So why should roboticists, who *can* achieve autonomy-as-independence with off-the-shelf technologies, seek even first-order autonomy-as-self-rule quite apart from complete autonomy-as-self-rule? Likewise, why should moral psychologists fret over whether human actions are ultimately rooted in complete autonomy-as-self-rule if the requirements of complete autonomy make it, if not unattainable, then at best *exotic*?[1]

Consider again the industrial robot toiling away in its well-defined factory environment. The distinction between this robot and the Mars Rovers or even the humble Roomba lies in the capacity of the latter to negotiate unpredictable environments. The behavior of the industrial robot is precisely scripted to within millimeter tolerances. Any change in the environment not anticipated by the industrial robot's programmers at best results in complete failure. At worst, it leads to the injury or death of human factory workers. The industrial robot is thus *brittle* with respect to its environment: It cannot suffer the least change in its environment without malfunction.

Unlike the industrial robot, the designers of the Mars Rovers and the Roomba have only the most general notions of what to anticipate in their robots' respective environments. Moreover, the problem of brittleness in robots is compounded by their own agency Thus Brooks saw quite clearly that robots by their very nature make changes to their immediate environment, whether by moving from one location to another or by directly manipulating objects. This creates something of a feedback loop of environmental dynamism, which is perhaps why brittleness in the smallest regard so quickly amplifies into catastrophe.

Yet the Mars Rovers and the Roomba are themselves brittle in comparison to a first-order

autonomous robot. Neither the Mars Rovers nor the Roomba has deliberative control over their action-directed deliberations, however rudimentary those action-directed deliberations may be. The Roomba cannot but return to its charging base, even if just how it navigates its way back is subject to measured responses to its environment. Similarly, the Mars Rovers cannot but traverse from one geological feature to another, even if how they negotiate impediments is not determined in advance by mission control. The first-order autonomous robot, on the other hand, has the capacity to assess, revise, reject, and even create new goals, giving it an altogether higher order of flexibility and adaptability over its more rudimentary predecessors.

If brittleness is the principle liability of agency, the utility of autonomy-as-self-rule lies in its inverse relationship with brittleness. That is, the greater the degree of autonomy-as-self-rule, the less brittle the agent insofar as the degree of the agent's autonomy is key to its adaptability.

Much the same is true of human agents. Our reflective, as opposed to reactionary, deliberative capacities have made us the technological savants of the world's ecosystem. Few environments on or off the Earth have proved thoroughly inaccessible to our adaptations. Indeed, our interest in completely autonomous robots is itself a manifestation of our own creative autonomy.

Works Cited

Brooks, R.A. 1999. "Intelligence Without Representation." In *Cambrian Intelligence: The Early History of the New AI.* Cambridge, Mass.: MIT Press. A Bradford Book.

Cummins, R. 1996. *Representations, Targets, and Attitudes.* Cambridge, Mass.: MIT Press. A Bradford Book.

Dennett, D. C. 1987. *The Intentional Stance.* Cambridge, Mass.: MIT Press. A Bradford Book.

Dworkin, G. 1988. *The Theory and Practice of Autonomy.* Cambridge: Cambridge University Press.

Frankfurt, H. G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68: 5-20.

Lewis, David. 1973. *Counterfactuals.* Cambridge, Mass.: Harvard University Press.

Swoyer, C. 1991. "Structural Representation and Surrogative Reasoning." *Synthese* 87: 449-508.

* I presented material from this paper at the North American Computing and Philosophy-2007 Conference held at Loyola University-Chicago. I'm grateful to conference participants for their puzzled looks and probing questions. I'm also indebted to Vere Chappell, Lynne Rudder Baker, Bruce Aune, and Rod Grupen, who contributed much to the development of these ideas.

1 I am grateful to Luciano Floridi for pointing out that the utility of autonomy-as-self-rule may not be entirely obvious.