What robots can, cannot, should do. What roboticists need.

Alberto Greco
University of Genoa

Paper presented at the Fourth EUCogIII Members Conference, Falmer/Brighton
"Social and Ethical Aspects of Cognitive Systems"
23-24 October 2013

Robots can be used for supplanting humans or to help humans. In the following, I discuss both perspectives.

Robots supplanting humans

There are two main kinds of objections against building/using robots that can *supplant* humans.

1: that is **impossible**, because humans are special and their capabilities could *never* be fully reproduced: this objection, however, is just a bet without no grounding. It is a bet against time, because time works against it, and since no one can demonstrate that there are human capabilities that *in principle* cannot be reproduced. All capabilities that can be translated into a language or into algorithms can be already fully reproduced by symbolic computations. And many competences that we don't know how to translate into algorithms can be modeled by dynamic or connectionist systems. However, this argument can lead to an infinite quarrel because, conversely, there is also no way to know when to stop and say that really *all* human capabilities have been reproduced.

2: that is **not admissible**, even if it were possible, because robots could not behave according to accepted ethics (like old Weizenbaum claim that an artificial psychotherapist would be "immoral"). But this argument will easily fall into the first one, because one is compelled to say what decision should be taken in the case that a robot fully behaving according to accepted ethics were ultimately built. A different argument for not accepting any artifact capable of supplanting humans is a problem of social inconveniences (e.g. they could expropriate humans of some jobs, or could not being accepted as interacting subjects), but this is like a "luddist" fight that is deemed to failure. Certainly, the argument can be rephrased to state that until robots will not exhibit human competences, in a *certain degree* (less than "all"), we should restrain from using them in place of humans. The most reasonable thing, then, is to find an agreement among roboticists (shared with the public opinion and stakeholders) about criteria for establishing what such "certain degree" means for acceptability.

Robots helping humans

Why robots helping humans should not be acceptable? For reasons similar to ones expressed above at point 2. I.e., because they lack some competencies that are judged essential to be really helpful. But new robots that possess such skills can always be built.

So the key question to be discussed is not about ethical questions in abstract, but about **what is acceptable** in different contexts, taking into account the features of present robots and also prospects that may reasonably be glimpsed. Robot features which have to be taken into account do not concern only limitations, but also aspects where robots overcome human capabilities, which is the real sense that makes worth using them (the main advantage of using a robot is to do what no human can do).

The roboticists community then needs guidelines and acceptability criteria. To determine what is acceptable, a careful evaluation of the potential trade-off between costs (including perhaps even prejudices and fears) and benefits is needed. To determine what are the benefits, a scale of values is needed, which go beyond the simplistic Asimov's law "not to harm humans": more subtle values

are to be put in such a scale, like not to affect some basic user's needs, ranging from physical survival to self-esteem or religious beliefs.

As an example, consider an assistive robot intended to help psychologically impaired people. Such a machine perhaps would not be presently easily accepted as a pleasant companion that can replace human companions ("presently": never say never, it could be like this in future). From this point of view, its implementation and use should not be recommended. But it could implement some psychological theories and, for example, could be capable of analyzing users' behavior, like their answers to questions, or measuring their reaction times, or recognize emotions, etc., better that any psychologist could do. Its acceptability could earn points. Such a robot could not be used, however, outside of any context. Even if it may be desirable to provide such a system the ability to make sense of contexts and situations on its own, in any case it must be clear that its aim is to provide a service: so it should be considered as a tool and the final control must be human. This kind of systems could be used, in sum, as a kind of a new generation of expert systems, but always under the supervision of psychologists, much like projective tests are used (in fact, it would be a sort of dynamic-interacting projective tool).