Self-Awareness implied in Human and Robot Intentional Action

Cristiano Castelfranchi and Rino Falcone

ISTC-CNR, Rome Italy
(cristiano.castelfranchi,rino.falcone)@istc.cnr.it

Abstract. In this paper we present a specific analysis of the consciousness theme. In particular, we are interested to identify how consciousness is relevant for the *intentional action* (both individual and social one) and in the construction of a *Self*. All these aspects are very important for understanding also how it is possible to build robots, or more in general autonomous artificial agents, able to realize deep and intelligent interactions.

Keywords: Consciousness, Intentional Action, Intentional Social Action.

1 Introduction

1.1 Our claims and the Ambiguity of "Consciousness"

In this paper our aim is to enunciate some clear theses about crucial forms of "consciousness" (a polysemic term[1]) and to exemplify and argue about them.

(1) A true *'intentional' action* requires a form of self-awareness, of meta-representation of mental states and of the self ("I").
(2) An *intentional 'social' action* and thus social interaction presupposes specific form of self-awareness and minds comparison.
(3) The *construction of a Self*, of a representation of the agent and its identity, obviously entails self-awareness.

Since robots should be able of real intentional (deliberated) actions (not just automatic behaviors), and of a real human-like interaction (at least with humans), and for this (and for other reasons) they have also to build and work on a *self-representation* and have some form of identity, thus also robots should have these forms of self-awareness.

1.2 Consciousness different from Subjective Experience, different from Mind

More precisely here we will not identify or reduce consciousness with/to the "subjective/phenomenological experience". Which is a problem of major interest which presupposes - in our view - a functionalist theory of the "body" and of "feeling" that we

[1] And "an elusive and controversial phenomenon", to use Chella and Manzotti' (6) words. And covering various and independent phenomena see also (15, 3).

consider realizable in artificial life. However, this is a problem we do not intend to focus on, although it is the prevalent and trendy interest, due to philosophy, "embodiment" frame in psychology, mental simulation view, etc.

1.3 Other form of Consciousness and Artificial Modeling

Let us put aside the problem of artificial modeling of phenomenal experience and *qualia*, and stress the priority of modeling other 'notions' of "consciousness":
- Self-awareness in terms of: our ***self-image*** (9, 12) and representation; an '***identity'*** (as cognitive representation of ourselves in social relations: our memberships, our features and qualities, our (comparative) evaluations; our biography and narrative, and those of our groups; our representations of future: impossible abandoned desires, plans and ambitions, hopes and expectations; our role duties, etc.; our exhibitions and maintenance of those representations, etc.); and our reasoning about it (which is a form and use of meta-cognition; see below); and our affective attachment to it.
- Meta-cognition processes: a) Representations and processing of our mental representations and processing. Like "believing to believe" or "believing to want/desire" or "desiring to believe" and so on. ***b)*** Mechanisms for internal self-monitoring and self-control. *Implicit meta-signals and meta-operations*: like "surprise" (signal of a cognitive mismatch), or "success feelings" and "failure feelings" (signals of matching or mismatching of world/belief and desires), etc. ... But, in particular and more relevant for consciousness, *explicit meta-representation and meta-reasoning* about mental states, and *meta-actions* upon them, in order to change our own goals and intentions, to change our own beliefs; being able to really have "will" and a "strength" of it; able to impose intentions to ourselves, and to exclude or block impulses or previous intentions; able to argue and persuade ourselves; and so on. And even *emotions about our emotions.*

2 Self-awareness as required in "intentional" agents

Let us focus in this paper on the relationship between a true "cognitively purposive", "intentional" behavior and forms of consciousness necessary for it.

2.1 I can, I know, I have to, ...

A really "intentional" action (1, 10), the building of a real "intention" to do something in order to achieve something, necessarily implies several aspects of Self Awareness; that is, not only goals about the world but goals about myself, beliefs about me, a representation of me (and of me in the world). Let's focus on specific and crucial components of such self-representation and evaluation. Any true "purposive behavior", driven by an anticipatory representation of the "goal" to be achieved, in a system endowed with multiple goals and thus with the need for choice, intrinsically requires some

form and components of consciousness, of self-awareness in the cognitive agent: reflexive beliefs and goals about (it)self. In particular, the "decision" or "deliberation" process, where the agent comes to formulate an "intention" to do a given action, implies some *beliefs* about the agent and its own mind. In order to decide to do something, to shape a goal as an "intention" of mine I have in fact to assume that such a goal is neither impossible nor independently realized, but that (a) it is "up to me", "it depends on me" (I have to conceive me as an "agent" with purposive effects on the world). Moreover, I have to believe that (b) "I know how" to pursue, realize that goal (I know the right plan, recipe, complex action, necessary for the specific contextual situation in which I will act); and (c) that "I'm able to", I have the skills for. Otherwise, I will not "do", decide "to do" that action, but I just "try", "attempt to", in order also "to see if.."; or I renounce, put aside that goal. It is not enough to just have that competence, we also need to *know* that we have it; the awareness of our powers, of our autonomy, is a condition for really having and exercising those powers, for being autonomous. These are crucial steps in the *goal processing* starting from desires and producing intentions. To "act" is to know ourselves; and it's *for* knowing ourselves, not only the world.

a) While *performing an action inthe world we see if we do really know "how" to achieve or do something, and if we have the right skills, and if our beliefs and predictions were right* or there was a mistake (we can meta-examine *our* reasoning to see what was wrong). There is in any action also an epistemic goal (not necessarily intentional and aware, but a function) to acquire knowledge about the world and us. Sometimes we just "try", "attempt"; that is, we perform the action with some subjective doubts about our competence, data, ability, or world conditions, thus we explicitly perform it for acquiring knowledge about: "to see if". Sometimes the trial is only for learning, is just an experiment.

b) By acting *we also send a signal about ourselves* (intentions, assumptions, values, abilities, powers, etc.) *not only to the others but to ourselves*. In order to see if we are able and succeed, but in order to "demonstrate" to ourselves that; and in order to show us who we are and confirm our self-image, our representation and mask of ourselves.

2.2 Consciousness for meta-decisions

Self-awareness is also crucial for some specific layer and kind of decision. In fact, we have also meta-decisions about deciding or not. We can for example deciding (choosing) of not deciding about something. This is not simply do not arrive to a decision, since the algorithm does not achieve a sufficient level of value discrepancy, or to suspend the decision because there is some important lacking data. This can just be a step, an exit of the decision procedure. We mean real meta-decision making based on evaluations of the pros and cons of deciding or not. For example, a classical case is deciding of do not decide in order to avoid the guilt, the responsibility of a wrong decision or of a taken risk, and self-blame. Notice that the risk of personally "taking" (not just "incurring in") a risk in a given decision is on both side on the choice: If we chose

A we “take” the A-risk, while choosing B we “take” B-risk. If we want “do not take a risk” we have to decide at a meta-level; to decide of do not decide between A or B. On one side there is risk-taking, on the other side rejection of taking that risks. Thus, in order to take such kind of meta-decision a serious form of consciousness is need; not only meta-representation, self-awareness of our cognitive decision processes, but also a self-representation about be object of possible reproach and self-blame, having the goal of not being blameworthy, the idea that we are “responsible” of the consequences of our decisions, etc.

Will our robots be able to decide of not decide? We think that this is a crucial ability for autonomous decision makers, especially in delegation relations, and for negotiating the level of delegation. Thus, we think that we have to build in our robots also these aspects of consciousness (8).

3 Consciousness in social interaction, social intentions

Let’s focus on our second thesis. Cognitive *sociality* necessarily implies some crucial and advanced aspects of Consciousness (and identity; distinction from the “other”). In fact - in order to interact and coordinate and cooperate with Y- we need to represent (ascribe) him mental state (what he believes and wants) but also his representation of our mind: what he believes that we believe and want; this probably in some symbolic format. Moreover, we have to *compare* these representations about our mental contents with our actual mental states (is Y wrong or right? Does he understand what we are doing, intending? Did I succeed in deceiving Y?); thus, reasonably we need a representation of our own mental states in the same format (*comparable*) for the representation of Y’s mental states. That is, we need to apply to ourselves the “mind reading” o “ascription” that we use socially (and we believe the other use on us).

Clearly this is an exceptional form (and original) of meta-cognition (for potential self-influencing, etc.,), of consciousness.

Can the robot be “wrong” in ascribing mental states to itself? Can be its reflexive mind-reading wrong? How does the robot “verify” or “read” in symbolic terms its mental contents and processes?

3.1 Self-awareness, Behavior Explanations, and Trust

A relevant issue about HRI -object of serious studies- is about the need of argumentations and explanations by the robot to the humans (or other robots) for clarifying the “reasons" of its choices: why it decided to do what it did, or why it did it that way. People claim - and we think that this is basically right- that this cognitive and communication ability be crucial for trusting robots. Can we trust an agent we do not understand what is doing and why? We would just to underline that this doesn’t mean that for trusting an (intelligent) system/process or an agent we have to understand how it

works; its “engineering”. This is frequently impossible, not necessary, and even wrong and counterproductive (as we argued in (5)). What is necessary is that we understand how it works for relying on it and using it; its affordance, dependability, and effectiveness. Do we know the real mechanism of automatic doors? Is this necessary for trusting them, and deciding to try to go through? No; but it is necessary that we understand that it open, surely and safely if/when a guy is approaching, and it remains open enough time for cross through it. So, we do not understand its “mechanisms” but we have a ‘mental model’ of how it works. Now it is true that, for relying on an “intelligent” and autonomous agent we depend on, we need such level of understanding: not only of its behavioral regularities or skills, but of its underlying perception, understanding, reasoning and deciding. We have to rely on its cognitive capabilities. For social interaction and coordination we need some for ‘mind reading’ or mind ascription to it. And even vice-versa: we want to understand if it understands what we know, want, our plan, etc. Now large part of this interpretation of its mind behind what is doing cannot be just inferential from its action, or just based on tacit behavioral communication. Sometime we are not able to understand what the other human is doing or why; it will be worst with “artificial” intelligences. So, it will be really important the capability of the robot to “explain” us what it is thinking, what is doing (which is a goal-notion, not an observational one) and why.

However, the problem is that in order to provide us such explanations about its behavior the robot has to describe its mental contents and processes. In other words, this will require an impressive meta-cognition, self-representation (“What **I** believe” “On which ground **I** believe so” “which was the process arriving to that decision, and on the basis of which elements” and so in deep). Take for example a simplified model of the step of goal processing from activation to a real Intention formulation. For explaining to others (and to myself!) why I decided in that way, I should be able to be aware and remind and communicate about at least 12 beliefs, and several active goals; and argue about the subjective value of those goals and my preference.

This crucial kind of trust – based on arguing and explaining about mental processes underlying the behavior – requires a remarkable “Self-Awareness”, consciousness, in a robot; but it is unavoidable for a really “social” relation.

A robot has to be able to answer to questions like: "*Who* has done that?" To answer that question the robot must have a representation of itself as agent distinguished from the "others"; but we wouldn’t call that "awareness". However, in order to answer questions like "*Why* did you do that" - as we said - it has to have a meta-representation on its own mental contents and process, and some meta-reasoning about its beliefs and goals and decisions. This definitely is a form of "consciousness".

Even more serious the problem for answering to: “Who has decided that?”; and being able to distinguish between just an order without alternatives, and its own decision; and be able to say: “I have decided so!”. But this requires a very analytic representation/interpretation (and memory) of its own mental processes out coming in that action; a very

conscious meta-observation. How could we trust a robot that is not aware of who has "decided" what it is doing? It is not able to be really consciously autonomous?

4 Conclusions

In this paper, we have primarily dealt with those forms of consciousness relating to intentional and cognitively purposive behaviors. We analyzed in some details different kinds of consciousness, in particular: self-awareness and meta-cognition processes. Let us conclude with a series of questions strictly linked with the problem of consciousness in Robots.

A) It could be useful to a Robot to have an "I"? To collaborate on a common plan, it is necessary to assign roles: "you" ("not me") do Action1; "I" do Action2. So -to work with- the Robot must have a notion of "me/I", or better a representation of the "self" as different from the representation of the "other". But is it really necessary to have a self-construct, a representation of the "self" as a form of awareness and identity? It would seem enough a merely procedural identification - not explicit – of the "self". It would be sufficient for the Robot to be programmed to perform all the actions mentioned in such a R24, and when the current plan / program specifies that there is now an action of R24, not R7 or H2, it "executes" that action: not a true "I", it seems enough a representation of the "self" in third person. The same may be true for plans to be executed alone over time: "I" act now but I also count on "mine" act next month to "complete" the plan I invested: I build a representation of "me" in the future and I rely on this. Likewise, to have projects and ambitions on the "future me": is it "I" or is it another? What kind of representation is necessary?

B) What is it for the unitarity of "I"? Not to cooperate with itself. Probably it is the result of other forms of consciousness such as the subjective experience and the distinction between an "inner" world and what I "feel" and an outside world that I feel and see. In fact if I (R24) have awareness of myself as R24 then I know that I must/can stay connected with the inner world of R24 and decide/ perceive/act "starting" from that inner world as a system of representation of reality. That is not identical to someone who perfectly reads the R24's internal world and acts accordingly, but has also its own dispositions, attitudes, decisions (we have to consider the influence of these attitudes albeit it tries to reduce it).

C) What does it mean that a Robot has real autonomous goals? An interesting problem regards the reduced motivational autonomy of existing machines. We can design Robot with "own" and autonomous goals, that is, Robot that decide to do or not to do things for their own goals (self-interested, self-motivated, not to be confused with *selfish*) and how to pursue them (sub-goals).[2]

[2] However, there is a limit to the autonomous nature of these (non-instrumental) end-goals that they pursue, as they are 'innate', as drawn. Different it would be the case for motivation not designed, but evolved through

There is a level where "autonomous goals/ends" means my own goals, internal goals; not remote controlled, no orders or injections. It does not change anything if it has given me (or in our case to the Robot) the natural selection or God (and how you know about this there is disagreement), or artificial selection or a designer, or my experience in the world. They are my *endogenous goals/ends* and they do not come all from the body. The problem of the agents' autonomy (from the world and from its stimuli, or from others) exists perfectly, even in disjointed minds.

Concluding, we can say that the types of consciousness discussed in this work (self-awareness, meta-cognition processes) if represented in the Robot would be not only interesting but useful. And it is not enough for a Robot to read what I have in mind (maybe to help me better) but also to compare what I have in mind with what it has in mind.

References

1. Bratman, M. Intention, plans, and practical reason. Cambridge, MA: Harvard University (1987).
2. Castelfranchi, C.. Self-awareness: Notes for a computational theory of intrapsychic social interaction. In G. Trautteur (ed.), *Consciousness*, Napoli, Bibliopolis, pp. 55-80. (1995)
3. Castelfranchi, C. Consciousness or consciousnesses? Modeling for disentangling. *International Journal of Machine Consciousness,* 2009, 1, n°2.
4. Castelfranchi C. (2012) "My mind". Reflexive sociality and its cognitive tools. In Paglieri (2012), John Bejiamins, pp. 125–150
5. Castelfranchi C., Falcone R., Trust Theory: A Socio-Cognitive and Computational Model, John Wiley and Sons, April 2010.
6. Chella A., Manzotti R. (eds.) Artificial Consciousness. Andrews UK Limited, 2013.
7. Dennett, D.C., (1971) "Intentional Systems", *The Journal of Philosophy*, Vol.68, No. 4, (25 February 1971), pp.87-106
8. Falcone R., Castelfranchi C. (2000), The Human in the Loop of a Delegated Agent: The Theory of Adjustable Social Autonomy, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, Special Issue on "Socially Intelligent Agents - the Human in the Loop, Volume 31, Number 5, September 2001, pp. 406-418
9. James, William. 1890. *The Principles of Psychology*. New York: Henry Holt.
10. Knobe. J. (2003). Intentional action and side effects in ordinary language. Analysis, 63
11. Kohlberg, Lawrence (1973). "The Claim to Moral Adequacy of a Highest Stage of Moral Judgment". *Journal of Philosophy*. The Journal of Philosophy, Vol. 70, No. 18. **70** (18): 630–646
12. Markus, Hazel. 1977. Self-schemata and Processing Information about the Self. *Journal of Personality and Social Psychology* 35 (2): 63–78.
13. Paglieri F. (ed). (2012) "Consciousness in Interaction: The role of the natural and social context in shaping consciousness". John Bejiamins, 2012
14. Piaget J., (1932), The Moral Judgment of the Child. Oxford, England: Harcourt, Brace.
15. Sloman A. (2010) An alternative to working on machine consciousness. *International Journal of Machine Consciousness* Vol. 2, No. 1 (2010) 118.

selection, learning. This would be an even more radical autonomy. This is another limit of 'naturalism', but it does not stop studying very important aspects of autonomy, self-regulation, even 'consciousness'.