Bye-Bye Agents? Not

Cristiano Castelfranchi • *Institute of Cognitive Science and Technology*

Is Carl Hewitt a good prophet in predicting a "Perfect Disruption: Causing the Paradigm Shift from Mental Agents to ORGs?"[1] I don't think so. The agents paradigm isn't a crisis yet and could play a more aggressive role not only in AI but in several sciences. I disagree that organizations of restricted generality (ORGs) are really an alternative to agents because they conceptually and computationally presuppose agents. But I take his paper as a wake-up call to the agent community.

Unfortunate Weaknesses

The agents paradigm is an attempt at a new scientific-technological paradigm, and it has rather invaded and transformed AI. But I agree with Hewitt that it has problems. In my view, the agents paradigm has five possible weaknesses:

1. semantic and conceptual carelessness — it lacks a convergent and unitary frame and formation;[2]
2. useless separation between autonomous agents in artificial life and adaptive systems, and robotics versus software agents and agent theory;
3. insufficient understanding of agents' scientific importance;
4. lack of convergent platforms and languages; and
5. lack of successful applications.

The AI agents community is aware of many of these deficiencies (see www.ofai.at/research/agents/conf/at2ai7; http://itmas2010.gti-ia.dsic.upv.es; and www.conferences.hu/AAMAS2009) but hasn't cared enough to repair them, particularly weaknesses 1 and 2.

As regards weakness 1, I strongly agree with Charles Petrie's call for an operational definition of agents in "No Science without Semantics."[3] In particular, I agree with his remarks about the unscientific vagueness and unfortunate consequences of terms such as agent, autonomy, and intelligence.

As for possible weakness 5, I'm unable to evaluate the impact and perspectives (see the work of Michael Luck and his colleagues for a good discussion[4]) or even the failure of agent technology.

It's reasonable to attribute some negative impact to factors 1, 2, and 4, but I wonder whether this is just the usual iterated and in a sense intrinsic declaration of failure of AI research and challenges: as long as they're explorative, they're AI; when they succeed, they're just IT. Thus, AI can only fail!

But this is an additional reason to rehabilitate AI's scientific nature, its ambitions — for instance, its role in computational (AI- and artificial-life-inspired) modeling for the biological, cognitive, sociological, and economic sciences.

This leads into my main obsession: that most of the community doesn't accept that agents aren't just an information technology; they're a way of thinking, a conceptual frame for modeling active, distributed, complex, and layered phenomena. Agents are a way of rethinking computing and knowledge in terms of interaction and social processing. They represent the accomplishment of a thinking path. Together with their networks and organizations, they're the most complete and suitable realization of a view of computers as *interactors* rather than simple *switches*, proposed by Joseph C.R. Licklider and Robert W. Taylor[5] following Wiener's view. Petrie's proposed operational definition of computers is necessarily interactional: it doesn't specify an agent architecture but only a very general mode of interaction that can be operationally verified.[6] Moreover, apart from the new perspective and conception of computing, agents are also a way of modeling the

 1089-7801/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

Editor's Note

This guest column is a response to Carl Hewitt's January/February 2009 guest column. I reserve this space for provocative columns and encourage readers to respond with either agreement or critiques, but especially the latter, of which this issue's column is an excellent example. Although both guest columnists disagree with each other, they agree with the same articles I've written. I invite other readers to disagree with me, always. — *Charles Petrie*

neural, cognitive, behavioral, and social sciences.

The AI community should realize this important potential contribution. Members of this community aren't just engineers, frequently importing some well conceived and effective formal theory from another discipline; they can export theoretical ideas and new conceptual apparatuses, not just application instruments. They can deeply change not only the neurocognitive and the social sciences with agents, but also contribute real control systems, neural networks, evolutionary algorithms, network dynamics, and so on.

ORGs Opposition: Agents Will (Hopefully) Survive

The general view of current disruption in Hewitt's article is interesting and basically correct, but I disagree with his diagnosis and prognosis for agents. I'm not fully able to evaluate the success, limits, and failures of agents from an engineering and technology viewpoint, but I think that Hewitt's perspective is too simplified and reductive.

He says, "A software agent is basically a mental agent adapted for software engineering," and that "a mental agent is defined behaviorally as cognitively operating in human-like fashion."[1] He sees them as individuals as opposed to organizations. This is a straw-man argument.

Hewitt's focus is on agents' specific cognitive architecture, but they're really a coordination model with many possible individual architectures. In fact, there's no clear distinction between agents, as practiced by the agents community, and the organizations that Hewitt describes.

There are four points in which I disagree with Hewitt.

Organizations Presuppose Agents

Agents aren't just a technology — even agent engineers should be proud of this. Agents are an epistemic and methodological framework for modeling theoretically, operationally, and experimentally complex problems characterized by the interaction of active distributed entities that have their own autonomous access to local and timely information and their own elaboration of those data building their local activity.

In this sense, they are autonomous; they get their own input and have their own internal state and purpose (not necessarily in a cognitive sense). They aren't simply stimulus-driven or responding to orders and executing fully specified instructions; they also work on the basis of their autonomous learning, exploration, elaboration, local conditions, and so on. They're driven by some internal representation (not necessarily explicit) and some internal aim (not necessarily explicit), that make it heuristic and useful to read them in the *intentional stance* (http://en.wikipedia.org/wiki/Intentional_stance).

Agents aren't isolated: they act in a common environment — that is, they affect the conditions and results of other agents — so they need some coordination.

Organizations actually are just specific and complex forms of coordination. Organization means "organization of or among ..." There are no real organizations without agents to organize, which are presupposed to be at least partially autonomous. Organizations presuppose agents and are grounded on them.

Agents achieve their collective and individual/local goals by interacting with other agents, with users, and with the environment and its coordination or knowledge artifacts. They might have individual interests (or represent specific interests) and thus find themselves in strategic situations.

A Scientific Frame and Tool

The "hopefully" in the section title isn't a corporate defense for me or the International Foundation for Multiagent Systems, it's a scientific "hope": the agent framework and technology is fundamental as a scientific frame and tool (notwithstanding the engineers and the repression of the AI revolutionary mission). Human, cognitive, and social sciences still need well-defined operational concepts and schemes for modeling their phenomena at the theoretical level. Moreover, they need models of (ways of conceptualizing) the proximate causal mechanisms and the hidden processes producing the observed phenomena, along with models of the dynamic processes of complex interactions, emergence, and self-organization. AI, artificial life, and autonomous agents and multi-agent systems can provide these models. Moreover, they can provide those sciences with not only conceptual but also experimental tools and new experimental data. As I said, social AI (and social artificial life) shouldn't just import concepts, theories, and phenomena from the cognitive, biological, and social studies and simply implement them or take them as inspiration for new technologies.

Of course, I do agree (this is an old AI issue) that because psychologically or socially inspired agents

can be very precious for science doesn't mean that engineers working on agents as technology for effective applications should necessarily model in such a way. Aircraft don't fly by fluttering! But, there are practical problems for which practical efficient solutions probably require some bio-inspired, socially inspired, and even human-like intelligence.

Agents Aren't BDIA Agents

Not all agents are "mental" (as Hewitt seems to assume). There are rule-based agents and agents based on simple neural nets and learning. All agents are reasonably goal-oriented — they have some function and some updated specific goal to achieve; but they aren't necessarily purposive or goal-driven — that is, they don't have an internal, explicit, anticipatory representation of the end-state to be realized, evaluating the world, selecting, guiding, and stopping the activity. Even the mental ones frequently don't really have beliefs, just data. Why identify agents as mental or BDIA (beliefs, desires, intentions, and affect) agents?

However, is Hewitt's prognosis — at least for this kind of agent — of marginalization grounded? As for engineering applications, I'm can't judge, but I don't think so. For sure, those agents and their platforms will prove valuable for the cognitive and social sciences, not only as experimental support but as operational models of internal and interactive processes (for the discussion on agents, see the work of Stan Franklin and Art Gasser,[7] Petrie,[6] and Michael Wooldridge and Nicholas R. Jennings.[8–9] See also the excited discussion on the agent list from August 2000 (subscribe by emailing agents@cs.umbc.edu) as well as my own work.[10–11]

Are Organizations not Agents?

Organizations are a fundamental new metaphor, and not just a metaphor but an abstraction, a conceptual schema for modern computing. They're much needed and promising, but are we sure that they're not agents at a different level of complexity and organization?

Agent is an abstract and functional notion; it shouldn't be applied to a specific physical support. There might be agents at different levels of complexity and granularity. If organizations aren't agents, how does it happen that Hewitt uses their activity concepts such as viewpoints, responsibility, agreements, work, authority, and so on? These are all typical features of social agents.

Organizations can be decomposed into specialized (a goal notion!) sub-organizations, whereas an individual (indivisible) agent can't be divided into subagents. But why should agents be individual? We can conceive complex agents (such as groups, teams, nations, and so on) made up of subagents. But there should be specific relationships not only among high-level activities, but between the higher-level complex goal or function and the lower-level missions and tasks — that is, goals to be achieved.

We shouldn't exclude the possibility of organizational "minds," that there are — in certain conditions — "minds" of a collective (AI should try to find the answer to this debatable issue).

The Real Problem

Much more alarming than Hewitt's diagnosis and prognosis are the discomforting data emerging from interviews with International Conference on Autonomous Agents and Multiagent Systems (AAMAS) participants that show there's not enough common view about the main issues and challenges of the AAMAS domain, no real common identity, no common history, no common core, no self-awareness; just a summation of small domains not understanding each other (weakness 1).[2] This isn't the right way to form young scholars, especially in a truly interdisciplinary domain.

Are AAMAS and related workshops real communities (implying a sense of identity: shared values and objectives, shared language, and common beliefs) or are they just a market: a place to go to expose my merchandise (and perhaps buy something), since there's some audience (clients), some credit, some reward (publication), and I go there to "sell" my "product" to some subgroup with very limited interests and topics.

Shouldn't agents (architecture and MAS) as a common view, a scientific project, and a general technological philosophy and approach be the central part of such a collective identity, common conception, and understanding? Doesn't this presuppose some conceptual, semantic clarification, and convergence and some greater effort for common platforms, languages, and issues? More intergroup discussions and training, and less sub-sub-sub-specialization?

I propose that the agent community should create a new initiative to define the various forms of agent technologies. It should be as inclusive as possible, but unless some technologies are excluded, nothing will have been defined. There should be a broad definition that represents the technologies that are included, and it should have some scientific value. We don't want to be in the business of defining "planet" or "agent."

We do have to be in the business of better defining what it is we do, how it's different, and showing that it matters. Unless we do a better job of defining our scientific contributions, agents will unfortunately continue to be sidelined as a legitimate scientific endeavor and community. This issue was discussed in part at the AAMAS 09 conference, and I hope that it will be covered in a more focused discussion at the next conference.

References

1. C. Hewitt, "Perfect Disruption: Causing the Paradigm Shift from Mental Agents to ORGs," *IEEE Internet Computing*, vol. 13, no. 1, 2009, pp. 90–93.
2. H. Cohelo, L. Antunes, and M. Luck, *Opinion Survey on Agents Community: Challenges and Trends*, tech. report, 2007.
3. C. Petrie, "No Science without Semantics," *IEEE Internet Computing*, vol. 11, no. 4, 2007, pp. 88, 86–87.
4. M. Luck et al., "Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing)," *AgentLink*, Sept. 2005.
5. J.C.R. Licklider and R.W. Taylor, "The Computer as a Communication Device," *Science and Technology*, Apr. 1968, pp. 21–41.
6. C. Petrie, "Agent-Based Engineering, the Web, and Intelligence," *IEEE Expert*, vol. 11, no.6, 1996, pp. 24–29.
7. S. Franklin and A. Gasser, "Is It an Agent or Just a Program? A Taxonomy for Autonomous Agents," *Proc. 3rd Int. Workshop Agent Theories, Architectures, and Languages* (ATAL 96), LNAI 1193, Springer, 1996, pp. 21–35.
8. M. Wooldridge and N.R. Jennings, "Intelligent Agents: Theory and Practice," *Knowledge Eng. Rev.*, vol. 10, no. 2, June 1995.
9. N.R. Jennings and M. Wooldridge, "Agent-Oriented Software Engineering," *Artificial Intelligence*, 2000, vol. 117, no. 2, pp. 277–296.
10. C. Castelfranchi, "To Be Or Not To Be An Agent," *Intelligent Agents III*, J. Muller, M. Wooldridge, and N. Jennings, eds., Springer, LNCS 1193, 1997, pp. 37–41.
11. C. Castelfranchi and R. Falcone, "From Automaticity to Autonomy: The Frontier of Artificial Agents," *Agent Autonomy*, H. Hexmoor, C. Castelfranchi, and R. Falcone, eds., Kluwer Publisher, 2003, pp. 103–136.

Cristiano Castelfranchi is the director of the Institute of Cognitive Sciences and Technologies of the Italian National Research Council and a full professor of cognitive sciences in the Department of Communication Science at the University of Siena. His research interests include the cognitive approach to communication (semantics and pragmatics); cognitive agent theory and architecture; multi-agent systems, agent-based social simulation, social cognition and emotions, and the cognitive foundations of complex social phenomena. Castelfranchi has a Laurea in Lettere from Roma La Sapienza. Contact him at cristiano.castelfranchi@istc.cnr.it.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.