CRISTIANO CASTELFRANCHI

Curriculum Vitae

Place/date of birth: Rome, 8 June 1944

E-mail: cristiano.castelfranchi@istc.cnr.it

Affiliation/address: Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Via S. Martino della Battaglia 44, 00185 Roma, Italy

Employment & Main Academic Duties

Researcher, Institute of Psychology, National Research Council (IP-CNR), 1971-2001
Full professor of General Psychology, University of Siena, 2001-present
Director, Institute for Cognitive Sciences and Technologies (ISTC-CNR), 2002-present
Full professor of Economical Psychology, LUISS University Rome, 2007-present
Full professor of Social Psychology, Uninettuno international telematic university, 2008-present

Education

1969: MA in Humanities with honours, University of Roma "La Sapienza", dissertation on "The representation of meaning"; supervisors R. Nencini (psychology), T. De Mauro (philosophy of language)
1973: 6-months at the Univ. of Vincennes (Paris VIII), linguistics with N. Ruwet, G. Fauconnier, J. Mehler
1974: 2-months at Berkeley University and Stanford University, semantics and AI with G. Lakoff, H. Clark, T. Winograd, R. Shank
2007: PhD honoris causa in Cognitive Science, University of Turin

Research Topics

- autonomy and goal-oriented behavior, with an emphasis on anticipatory action-control
- cognitive agent theory and architecture, with special focus on goals and their dynamics
- cognitive foundations of social phenomena (trust, power, cooperation, norms, institutions, etc.)
- cognitive approach to communication (semantics and pragmatics)
- social cognition and emotions, with an emphasis on the cognitive anatomy of complex emotional states
- multi-agent systems and social simulation, integrating cognitive, social and computer science

Scientific Bio

Castelfranchi started collaborating with the Institute of Psychology of the National Research Council (IP-CNR, now ISTC-CNR) already during the final years of his undergraduate degree, and he became a permanent researcher there in 1971. In those early years his interests were divided between two main topics: the *pragmatics* and *semantics* of natural language (with D. Parisi), with an emphasis on generative linguistics and how explicit representations of mental states (goals and beliefs) determine language understanding, and the *analysis and prevention of mental disease* with non-constrictive methods (with F. Basaglia and R. Misiti), which was instrumental to the legal reform of the national psychiatric system in Italy in 1978. His theoretical approach was focused since the onset on defining *an operational notion of goal*, partially inspired by progresses in cybernetics and control theory, in sharp contrast with the more vague and affective notion of "motivation", traditionally employed in cognitive and social psychology at that time. Defining goals as *anticipatory representations of world-states capable of guiding the agent's behaviour* proved extremely helpful in understanding language, and it became soon clear to Castelfranchi that this notion was equally crucial in understanding social phenomena in general, as well as in elucidating the cognitive structure of complex emotions. This led Castelfranchi to initiate mayor research programmes in the study of *social norms* (with R. Conte), *trust* (with R. Falcone), and *emotions* (with M. Miceli and I. Poggi), which resulted in significant breakthroughs in each of these areas. In parallel, the emphasis on operational, theory-driven conceptual notions, as opposed to the traditional ill-defined, data-driven constructs of psychology, fuelled Castelfranchi's exchanges and collaborations with people working in computer science, including some of the founding figures of modern AI (such as T. Winograd and R. Shank at Stanford in the early '70s). In particular, in the '80s Castelfranchi worked intensively on *computational linguistics* (with O. Stock and D.

Parisi), and later on became one of the key figures in the creation and consolidation of the *multi-agent systems* approach to Distributed Artificial Intelligence. Castelfranchi's interest in the autonomous agents paradigm and in the use of agent-based simulation to analyze social phenomena was motivated by the strengths and weaknesses he perceived in that seminal area: agent-based models naturally stressed the autonomous and cognitive nature of the agent's architecture, using notions that were at the same time clearly defined and operational, but also naive and not enough informed by psychological and social science. Castelfranchi clearly saw the need to provide a robust theoretical foundation for some key notions in multi-agent systems (such as power, dependence, norms, and commitments), thus playing a unique role in shaping since the onset this now thriving research community. At the same time, Castelfranchi clearly recognized the potential relevance of this methodology for the understanding of human cognition and society: in a series of influential research projects and papers, Castelfranchi systematically investigated the *cognitive mediators of social phenomena*, using the agent-based approach as a conceptual tool (a theory rather than a technology) to analyze society as emerging from the interaction of cognitive agents, and cognition as being shaped by social interaction – a framework that later gained relevance not only in computer science, but also in social psychology, economics, and philosophy. All these different and complementary research approaches came to full fruition in the last decade of Castelfranchi's scientific activity: the *role of society in shaping cognitive processes* and the *centrality of action-control* became mainstream in cognitive science and neuropsychology, vindicating Castelfranchi's idea (which he championed since the '70s) that *cognition is essentially for action*, rather than the mere information-processing assumed by traditional approaches. This gave Castelfranchi a central role, over the last decade, in studying some essential features of what are nowadays known as *cognitive systems* (http://cordis.europa.eu/fp7/ict/cognition/home_en.html), especially on anticipatory mechanisms, goal-oriented action-control, and autonomous behaviour.

Responsibilities

Director, Institute for Cognitive Technologies of the National Research Council (ISTC-CNR, 2002-present)
Head, Division of "Artificial Intelligence, Cognitive and Interaction Modeling", and Division of "Social Psychology" at the Institute of Psychology of the CNR (1971-2001)
Founder and director, Goal-Oriented Agents Laboratory (GOAL) at the ISTC-CNR (2004-present)
Coordinator, National Finalized Project of the CNR on "Prevention of Mental Diseases" (1979-82)
Reviewer for research projects with the European Community (FP6 and FP7), the European Science Foundation (ESF), the Netherlands Organisation for Scientific Research (NWO), the Spanish Ministry of Science and Innovation (MICINN), the Portuguese National Science Foundation (FTC), the Bulgarian Academy of Science (BAS), and the Italian Ministry for Research and Higher Education (MIUR).

Main awards and honours

1991: Honorary fellow of the Italian Society for Behavioural and Cognitive Therapy (SITCC)
2003: ECCAI fellow, for pioneering work in Artificial Intelligence
2007: Mind & Brain prize, Univ. of Turin (co-awarded with M. Tomasello), for pioneering the integration of psychology in cognitive science and breakthroughs on autonomous agents and their interactions

Current/submitted research grants (funding ID) *(approximate amounts for Castelfranchi's lab only)*

- EC-FP7/STREP, submitted, *Flexible Recombination of Embodied Experiences in Different Domains* (FREEDOM), co-investigator (15% research time), total funds: 680.000 €
- EC-FP7/STREP, 2011-2014, *Goal-directed, Adaptive Builder Robots* (Goal-Leaders), co-investigator (5% research time), total funds: 475.000 €
- EC-FP7/CA, 2011-2014, *European Network for Social Intelligence* (SINTELNET), leader WG on Social Epistemology (5% research time), total funds: 65.000 €
- EC-FP7/STREP, 2009-2011, *Humanoids that Learn Socio-Communicative Skills by Observation* (HUMANOBS), co-investigator (5% research time), total funds: 370.000 €
- EC-FP7/COST, 2008-2012, *Agreement Technologies* (AT), leader WG on Norms (5% research time), total funds: 80.000 €

Past research grants *(selection)*

- Bilateral project (Italy/USA), 1992-93, *Dependency and Psychological Wellbeing*, with J. Brehm
- EC-ESPRIT, 1994-97, working group on *Modelling Agents for Information Technologies* (ModelAge)
- Bilateral project (Italy/Bulgaria), 1995-97 & 1998-2000, Cognitive Modeling, with B. Kokinov

- EC-FP5 & FP6/CA, 1998-2006, *European Network of Excellence on Agent Research*: AgentLink I (1998-99), AgentLink II (2000-04), and AgentLink III (2004-06)
- ESF/EUROCORES, 2003-07, PI of *The Cultural Self-organisation of Cognitive Grammar* (CUSCOG), OMLL programme (*The Origin of Man, Language and Languages*), with L. Steels, P. Dominey
- EC-FP6/STREP, 2004-07, *From Reactive to Anticipatory Cognitive Embodied Systems* (MindRACES)
- EC-FP6/NoA, 2004-07, *Human-Machine Interaction Network on Emotion* (HUMAINE)
- ESF/EUROCORES, 2006-09, PL of *Consciousness in Interaction: The Role of the Natural and Social Environment in Shaping Consciousness* (CONTACT), CNCC programme (*Consciousness in Natural and Cultural Context*), with N. Frijda, A. Clark, S. Hurley, T. Metzinger
- ESF/EUROCORES, 2007-10, PI of *The Social and Mental Dynamics of Cooperation* (SOCCOP), TECT programme (*The Evolution of Cooperation and Trade*), with S. Bowles, E. Fehr, H. Gintis, R. Mace

Top 10 Publications, 2001-2011

1. Castelfranchi, C. (2001). The theory of social functions. Challenges for multi-agent-based social simulation and multi-agent learning. *Cognitive Systems Research* 2, 5-38. Cit.: **63** (5.7/year)
2. Castelfranchi, C., Tan, Y.-H. (2002). The role of trust and deception in virtual societies. *International Journal of Electronic Commerce* 6 (3), 55-70. Cit.: **98** (8.9/year)
3. Castelfranchi, C. (2003). The micro-macro constitution of power. *ProtoSociology, An International Journal of Interdisciplinary Research* 18-19, 208-269. Cit.: **70** (7.8/year)
4. Castelfranchi, C. (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic* 1, 47-92. Cit.: **58** (6.8/year)
5. Castelfranchi, C., Falcone, R., Pezzulo, G. (2003). Trust in information sources as a source for trust: a fuzzy approach. In: *Proceedings of AAMAS'03*, New York. ACM, pp. 89-96. Cit.: **70** (7.8/year)
6. Castelfranchi, C. (2005). Mind as an anticipatory device: For a theory of expectations. In: M. De Gregorio, V. Di Maio, M. Frucci, C. Musio (eds.), *Proceedings of Brain, Vision, and Artificial Intelligence*. Berlin: Springer, pp. 258-276. Cit.: **29** (4.1/year)
7. Castelfranchi, C., Paglieri, F. (2007). The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* 155*,* 237-263. Cit.: **57** (11.8/year)
8. Pezzulo, G., Castelfranchi, C. (2007). The symbol detachment problem. Cognitive Processing 8 (2), 115-131. Cit.: **24** (4.8/year)
9. Miceli, M., Castelfranchi, C. (2007). The envious mind. *Cognition and Emotion* 22, 449-79. Cit.: **23** (4.6/year)
10. Pezzulo, G., Castelfranchi, C. (2009). Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological Research* 73, 559-577. Cit.: **9**[1] (3.0/year)

Book chapters *(selection, 2001-2011)*

11. Falcone, R., Castelfranchi, C. (2001). Social trust: A cognitive approach. In: C. Castelfranchi, Y.-H. Tan (eds.), *Trust and Deception in Virtual Societies*. Dordrecht: Kluwer, pp. 55-90. Cit.: **249**[2] (23.7/year)
12. Falcone, R., Castelfranchi, C. (2001). The socio-cognitive dynamics of trust: Does trust create trust? In: R. Falcone, M. Singh, Y.-H. Tan (eds.), *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives*. Berlin: Springer, pp. 55-72. Cit.: **63** (5.7/year)
13. Castelfranchi, C., Falcone, R. (2003). From automaticity to autonomy. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), *Agent Autonomy*. Dordrecht: Kluwer, 103-136. Cit: **29** (3.2/year)
14. Castelfranchi, C. (2003). For a "cognitive program". Explicit mental representations for *Homo oeconomicus* (the case of trust). In: N. Dimitri, M. Basili, I. Gilboa (eds), *Cognitive Processes and Economic Behaviour*. London: Routledge, pp. 168-208. Cit.: **8** (1.0/year)
15. Castelfranchi, C. (2003). Grounding we-intentions in individual social attitudes. In. M. Sintonen, P. Ylikoski, K. Miller (eds.), *Realism in Action: Essays in the Philosophy of Social Sciences*. Dordrecht: Kluwer (Synthese Library). Cit.: **11** (1.2/year)
16. Castelfranchi, C. (2004). Trust mediation in knowledge management and sharing. In: C. Jensen, S. Poslad, T. Dimitrakos (eds.), *Trust Management*. Berlin: Springer, pp. 304-318. Cit.: **20** (2.5/year)

[1] Quoted (in Cisek & Kalaska 2010) as one of the most promising hypothesis on the evolution of anticipatory action-control in the *Annual Review of Neuroscience* (impact factor 2009 = 24.822).

[2] This article is often misquoted (111 out of 249 citations) as having Castelfranchi as the first author.

17. Paglieri, F., Castelfranchi, C. (2005). Revising beliefs through arguments: Bridging the gap between belief revision and argumentation in MAS”. In: I. Rahwan, P. Moratïs, C. Reed (eds.), *Argumentation in Multi-Agent Systems* Berlin: Springer, pp. 78-94. Cit.: **24** (3.4/year)
18. Castelfranchi, C. (2006). Cognitive architecture and contents for social structure and interactions. In: R. Sun (ed.), *Cognition and Multi-Agent Interaction*. Cambridge: CUP, pp. 355-390. Cit.: **5** (1.0/year)
19. Castelfranchi, C., Falcone, R., Marzo, F. (2006). Being trusted in a social network: Trust as relational capital. In: K. Stølen, W. Winsborough, F. Martinelli, F. Massacci (eds.), *Trust Management*. Berlin: Springer, pp. 19-32. Cit.: **15** (2.5/year)
20. Paglieri, F., Castelfranchi, C. (2006). The Toulmin test: Framing argumentation within belief revision theories. In: D. Hitchcock, B. Verheij (eds.), *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation*. Berlin: Springer, pp. 359-377. Cit.: **15** (2.5/year)
21. Tummolini, L., Castelfranchi, C. (2007). Trace signals: The meanings of stigmergy. In: D. Weyns, H. Van Parunak, F. Michel (eds.), *Environments for MAS*. Berlin. Springer, pp. 141-156 Cit.: **13** (2.6/year)

Monographs & Edited Volumes *(selection, 2001-2011)*

22. Castelfranchi, C., Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Approach.* Chichester: Wiley (370 pp.). Cit.: **11** (4.8/year)
23. Decker, K., Sichman, J., Sierra, C., Castelfranchi, C. (eds.) (2009). *Proceeding of 8th International Conference on Autonomous Agents and Multiagent Systems*. New York: ACM.
24. H. Hexmoor, C. Castelfranchi, R. Falcone (eds.) (2003). *Agent Autonomy*. Dordrecht: Kluwer (288 pp.).
25. Castelfranchi, C., Lesperance, Y. (eds.) (2001). *Intelligent Agents: Theories, Architectures, Languages*. Berlin: Springer (386 pp.).
26. Castelfranchi, C., Tan, Y.-H. (eds.) (2001). *Trust and Deception in Virtual Societies*. Dordrecht: Kluwer (292 pp.). Cit.: **92** (8.4/year)

Overall citation analysis *(data from PublishOrPerish, consulted on April 1, 2011)*

- **Overall**: 412 contributions, 7582 citations (per year: 176), h-index=**41**, *g-index=79*
- **2001-2011**: 190 contributions, 2239 citations (per year: 204), h-index=24, g-index=42
- Solid research track across his whole career, still very active and promising in the last decade: e.g. Castelfranchi’s number of citations per year is steadily growing, as well as the volume of his publications.

Invited Presentations: Conference as Keynote and Advanced Schools as Faculty *(selection, 2001-2011)*

2011 3rd International Conf. on *Agents and Artificial Intelligence* (ICAART), Rome
2009 Brazilian International Meeting on *Cognitive Science* (EBICC), Campinas, with M. Dascal, C. Sinha
2008 Advanced school *Minds & Societies*, Montreal, with D. Dennett, L. Sanger, D. Roy, J. Byrne
2008 Advanced school *Social Norms*, San Sebastian, with J. Elster, D. Sperber, H. Gintis, C. Bicchieri
2008 Advanced school *Social Cognition and Social Narrative*, San Marino, with S. Gallagher, S. Stich
2007 Europ. Conf. of *Cognitive Science* (EuroCogSci), Delphi, with G. Gigerenzer, R. Gallistel, M. Boden
2007 Workshop *Theory of Mind*, Marina del Rey, US, with A. Damasio, M. Iacoboni, A. Goldman
2007 Workshop *Expectation and Surprise*, Singapore, with A. Ortony, R. Reisenzein, J. Schmidhuber
2006 Workshop *Minds in Interaction* at the Netherlands Institute for Advanced Study (NIAS), Wassenaar
2005 1st Symposium *Artificial Economics*, Lille, with R. Axtell
2004 Workshop *Changing Minds*, Institute for Logic, Language and Computation (ILLC), Amsterdam
2002 Advanced school *Cognitive Processes & Economic Behaviour*, Siena, with M. Bacharach, P. Suppes

Organization of International Conferences & Committee Memberships *(selection, 2001-2011)*

2003-11 PC Member, *European Conference of Cognitive Science* (EuroCogSci 2003, Osnabruck; 2007, Delphi; 2011, Sofia)
2010 Co-Chair, workshop on *Norm Compliance*, EUI, Fiesole (with R. Mace, D. Sperber, G. Sartor)
2009 General Chair, *8th International Joint Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS’09, Budapest)
2005 ESF Exploratory Workshop on *Understanding the Dynamics of Knowledge*, Siena (with P. Gardenfors, J. van Benthem, R. French, L. Robinson)
2004 4th conference on *Collective Intentionality*, Siena (with M. Tomasello, J. Heinrich, M. Bratman)
2002 Program Chair, *1st International Joint Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS’02, Bologna)

Present PC Member in more than 30 international workshops and conference on innovative approaches and frontiers topics (20 of them on EasyChair)

Prizes, Awards and Academy Memberships *(selection, 2001-2011)*
Mind & Brain prize 2007 (with M. Tomasello) and *PhD Honoris Causa in Cognitive Science*, Univ. of Turin
Fellow, European Coordinating Committee for Artificial Intelligence (ECCAI, 2003)
Member, European Steering Committee for Cognitive Science, coordinated by S. Vosniadou (2003-present)
Co-founder and board member, Italian Association for Cognitive Science (AISC, 2002-present)
Service Award, Association for Computing Machinery (ACM, 2002)
Emeritus Member, Board of Directors of the International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS; 2002-present)

Editorial Boards
EB member of *Mind & Society* (with K. Arrow, J. Elster, D. Sperber, A. Goldman, E. Shafir, P. Slovic, C. Camerer, A. Cicourel, J. McClelland, J. Fodor, J. Searle, K. Binmore, P. Hogarth), *Cognitive Science Quarterly* (with D. Kayser, W. Kintsch, K. Stenning), *J. of Autonomous Agents and MAS* (with J. Rosenschein, M. Wooldridge, C. Sierra), *J. of Artificial Societies and Social Simulation* (with R. Axtell, N. Gilbert, R. Hegselmann), *Argument & Computation* (with P. Dunne, P. McBurney, S. Parsons, D. Walton), *Journal of Mind Theory* (with B. Baars, P. Gardenfors, J. McCarthy, A. Sloman) and the MIT *CogNet*
Co-founder and EB member, Italian journal of cognitive science and Artificial Intelligence, *Sistemi Intelligenti* (1989-present, with D. Parisi, V. Gallese, C. Umiltà, M. Piattelli Palmarini, P. Legrenzi)