Introduction

Getting to know the pupils of a true master speaks volumes about the mindset and skills of their mentor. So, before discussing Cristiano Castelfranchi’s intellectual odyssey (still very much in progress, we are happy to say), let us mention a few things about people that had the good luck of working with him, at various stages of their own education and/or career. In place of the somewhat cumbersome and pompous label of “pupils of Castelfranchi”, we will christen them (us) as “Castelfrankians”.[1] Here are a few points worth mentioning about this odd group of people:

- *Castelfrankians live among us!* Over the years, Castelfranchi managed to spawn a veritable host of students, colleagues, collaborators, and the like. So, be warned: Castelfrankians are many and multifarious (see next point), and you might be sitting next to one of them – come to think of it, if you are reading this, you are probably a Castelfrankian yourself!
- *Castelfrankians are not like him!* When Castelfranchi feels the urge to look at himself, he gets in front of a mirror; when he wishes to converse with himself, he does just that – often loudly, truth be told. What he does not, and never did, is imposing his ways to his pupils. Quite the opposite: autonomy and argumentativeness are the two features that Castelfranchi prizes above all in his interlocutors. As a result of that, not a single Castelfrankian does research as Castelfranchi does – not in the sense that they are not as good as him (some are, in their own ways, which is indeed remarkable), but in the sense that they do research differently from him, sometimes even drastically so, in more or less open conflict with his ideas and methods. Oddly, he does not mind that – in fact, the more confrontational former pupils turn out to be, the prouder they make him.
- *Castelfrankians are always more conservative than Castelfranchi!* This is one of three things that all Castelfrankians share (the other two follow below): no matter how hard they try, they will never manage to outdo their mentor

[1] Let us take this opportunity to set the phonetic record straight: the “chi” in “Castelfranchi” is correctly pronounced \’ki\, as in “kitchen”, and not \’chi\, as in “children”. Thus, to avoid the misspelling and mispronunciation that often plagued their mentor, Castelfrankians have opted for using the *k* to remove any ambiguity in their name.

in terms of revolutionary ideas, unconventional style, and sheer (but kind) disregard for the rituals and etiquette of academia. This is especially disconcerting, and at times downright depressing, for relatively young Castelfrankians: they would love nothing more than play the role of the innovative genius who rebels against the oppressive and outdated rule of their academic father figure, yet this pleasure is forever denied them. On the contrary, already at an early age they must learn to embrace the opposite attitude, counseling prudence and greater academic decorum to the scientific daredevil they happen to be working with. Castelfranchi, who probably never watched a single episode of *Star Trek*, is nevertheless thoroughly committed "to boldly go where no man has gone before", as per the motto of the famous spaceship. In doing just that, his style cannot be imitated, his momentum cannot be matched: this makes for a wonderful scientific ride, but forces Castelfrankians to play the part of the reactionaries.

- *Calstelfrankians are disciplinary nomads!* This is one of the most lasting effects of working with Castelfranchi: after some time, you discover that you are no longer able to conceive your entire scientific life within just one discipline, not even if doing so would ensure better chances of "academic survival" – as it is often the case, alas! Even more drastically, Castelfrankians end up looking at disciplines as mere instruments: they learn from their mentor that all that really matters in research are problems, that is, phenomena in need of explaining, and what tools one uses to do the explaining are irrelevant, as long as they are well suited for the task and applied with scientific rigor. Thus, Castelfrankians do not feel a sense of belonging to any particular "disciplinary church", and they fail to understand what is all the fuss about that. It is not as if they were jacks of all trades (they are not), it is just that they feel the need to specialize on specific problems, rather than on specific methods. Sometimes, this makes for an awkward living in today academia, where interdisciplinarity is regularly paid lip service, but rarely helps anyone getting tenured. Yet, the benefits of a life full of problems outdo the discomforts of a career without disciplines – see next point.
- *Castelfrankians have fun!* Happiness, for Castelfranchi, is in the pursuit of one's goals, not in their attainment. It is no happenstance that one of his favorite quotes is the well-known quip by Richard Feynman: "Science is like sex: sometimes something useful comes out, but that is not the reason we are doing it"[2]. As a modern day Sherlock Holmes, Castelfranchi feels alive, professionally speaking, only when the game is afoot, that is, when there is some intellectual puzzle occupying all of his attention and strenuously evading his attempts to solve it. The end of problems to struggle with would be for him nothing short than intellectual death. Conversely, a scientific problem is not a personal threat, in the sense that failing to understand something is, for Castelfranchi, the beginning of an exciting adventure, instead of something to be feared or avoided – not because one is necessarily sure of finding the right solution (or any solution), but because looking hard for it will be rewarding whatever the outcome, and well beyond it.

[2] Often quoted in a slightly different version: "Physics is like sex: sure, it may give some practical results, but that's not why we do it". The meaning, though, is the same.

> This playful attitude is something all true Castelfrankians share with their mentor: in fact, such an attitude is the only strict requirement for working with him – if you do not have it, you cannot fake it. Unless you love struggling with problems, a day with Castelfranchi is a day in hell! Even at the apex of some grand scientific revelation, or in the middle of some well-deserved public celebration of his achievements, he will always manage to remind you (and himself) that the problems yet to be solved vastly outnumber those we already dealt with. What is worse, he will say that with a twinkle in his eyes and a smile on his lips.

This brief sketch of the Castelfrankians might strike some reader as too quixotic to do justice to their mentor. Indeed, those who do not know Castelfranchi might very well picture him in the act of appropriating Carl Jung's famous (alleged) quote, and exclaiming: "Thank God I'm Castelfranchi and not a Castelfrankian!"[3]. But for those who do know him, it will be easier to imagine Castelfranchi recognizing himself in this description of his former pupils, which of course was meant to be humorous, yet not too far off the mark. And, just to avoid offending anyone, let us be clear: we, the editors of this volume, are Castelfrankians of the purest breed, and proud to be! With all our quirks and strange habits, we would not want to do research in any other way, and we feel nothing but honored to trace back part of our scientific DNA to Cristiano.

More to the point, the broad scope and problem-oriented nature of Castelfranchi's research is fully reflected in the thirty-nine essays gathered in this volume, to honor his incredible career. Some of them have been written by true Castelfrankians, but the vast majority is authored by many of the scholars who happened to collaborate with him over the years, and/or have been influenced by his ideas. Some others were invited but could not participate in this volume, for a variety of reasons: however, as editors, we were taken aback by how few people declined our invitation, always expressing genuine regret and mentioning truly insurmountable difficulties. For any one who declined, at least five more were eager to accept. This gave us the measure of how deep Castelfranchi's scientific influence and personal kinship is felt in all the academic circles where he ventured during his career – which are indeed many, as this volume demonstrates.

In spite of this diversity, the contents of this book are remarkably coherent, and even systematic. They exhibit the same kind of coherence which is found in the solar system, where all planets differ in size, mass, structure, motion, and much more, yet they all revolve around the same star. For the papers in this collection, this center of gravity is the notion of *goal*. We are reminded of its centrality not only by the very title of this volume, but also by Castelfranchi's own contribution, at the very end of it. In that final chapter, he discusses goals as "the true center of cognition", around which all the rest is built. Incidentally, taking that perspective allows him to recap most of his scientific endeavors, spanning decades of research, since he consistently devoted his distinguished

[3] The phrase "Thank God I'm Jung and not a Jungian!" is attributed to Jung in Yandell, J. (1978). The imitation of Jung. An exploration of the meaning of "Jungian". *Spring*, 54–76 (in particular, on p. 57).

career to *understand cognition as goal-directed*. In his own words, "it's all about goals: action is for goals (and goals are for potential actions), knowledge is for goals, intelligence is for goals (solving problems via mental representations), sociality is for goals and goal-based, [and] emotions are goal-centered".

Around this pivotal and multifaceted notion of goals, of which Castelfranchi has been one of the foremost pioneers, all contributions to this volume orbit. Part I, "Representation, action and cognition", provides some essential groundwork to the whole edifice: goals are analyzed as anticipatory representations, whose main function is precisely to bridge the gap between cognition and action – more exactly, to make sure that cognition is *for* action, rather than disjointed from it. Giovanni Pezzulo (chapter 1) outlines Castelfranchi's project of re-founding cognitivism on the cybernetic idea of goal-directed action, spelling out its implications for our understanding of the mind and its connections with many recent developments in cognitive science. Martin Butz (chapter 2) discusses how anticipatory agents form spatial representations for flexible, goal-directed decision making and behavioral control, thus laying the foundations for the development and grounding of higher-level, symbolic cognition and abstract thought. The issue of development is then taken up by Marco Gori (chapter 3), who considers the implications of stage-based cognitive development for the next generation of learning systems and models in Artificial Intelligence.

Part II, "Reasons, reasoning and rationality", discusses how Castelfranchi's goal-centered view of cognition impacts on issues of reasoning and rationality, providing a garden variety of test cases and applications. Aldo Franco Dragoni (chapter 4) details a computational structure for representing recursive mental states, intended to constitute the semantic level of a formal language to deal with cognitive dynamics. In a similar vein, Mehdi Dastani and Leendert van der Torre (chapter 5) revisits the well-known BOID architecture (beliefs, obligations, intentions, desires), to demonstrate how it can handle goal generation in agent systems. Fabio Paglieri (chapter 6) focuses on beliefs, using Peirce's notion of "the irritation of doubt" to highlight the goal-centered nature of belief dynamics, and to uncover a close kinship between pragmatism and goal theory. David Godden (chapter 7) shows how the psychology of belief, and the role goals play in it, should inform process-based accounts of argumentation, at the same time posing two problems for assessing the rationality of arguments. Paolo Legrenzi and Alessandra Jacomuzzi (chapter 8) discuss the goals of analogies, describing the mechanisms of transfer of the solution from one problem to another, as well as the main cognitive and computational theories of analogical reasoning. The attention shifts towards decision-making with Nicola Dimitri (chapter 9), who points out a behavioral duality in intertemporal choice regarding losses and gains, and suggests a more complex role of delay tolerance in explaining such choices. Time and decision are also central in Maury Silver's contribution (chapter 10), where he fleshes out an intriguing account of self-deception as procrastinating further investigation on current evidence, and then proceeds to illustrate the import of this view for the much debated notion of self-deception.

Part III, “Emotion and motivation”, takes advantage of Castelfranchi’s rich work on emotions as goal-centered constructs, to develop it in further details and unexpected directions. Rainer Reisenzein (chapter 11) contrasts the belief-desire compound theory of emotion, that he attributes to Castelfranchi, with his own causal feeling theory, and finds the former wanting in its ability to provide a convincing characterization of emotional states. Giorgio Coricelli and Mateus Joffily (chapter 12) focus on the neural correlates and the role of cognitive-based emotions in decision-making, arguing that cognitive processes, such as counterfactual thinking and social comparison, elicit a specific class of emotions, of which regret and envy constitute paradigmatic examples. Isabella Poggi and Francesca D’Errico (chapter 13) analyze pride in connection to the goals of power, image and self-image, thus distinguishing three types of pride (dignity, superiority, and arrogance pride), discussing their functions as displays of dominance, and connecting them to bodily expression in political debates. Francesco Mancini and Amelia Gangemi (chapter 14) illustrate the clinical implications of a goal-centered approach to cognition and emotions, in relation to depressive reaction and its two main paradoxes, pessimistic fixation (why do depressed people continue to dwell on what they believe to be unattainable and/or lost forever?) and lack of motivation (why do depressed people lose interest in alternative goals, instead of trying to compensate for the loss they suffered?). Emiliano Lorini (chapter 15) revisits and extends his own analysis of expectation and expectation-based emotions (e.g., hope and fear), which he originally co-authored with Castelfranchi, by distinguishing between goal value and belief strength in the anatomy of expectations, thus providing a formal analysis of the intensity of hope and fear. Samuel Bowles and Herbert Gintis (chapter 16), building on Castelfranchi’s work on the relationship between norms and emotions, model the process by which social emotions (e.g., shame, guilt, pride) can positively affect social interaction, by favoring high levels of cooperation with minimal levels of costly punishment, and discuss under what conditions such emotions might have evolved.

This also serves to introduce the rest of the volume, which describes in great details how Castelfranchi’s goal-centered view of cognition radically transforms and deepens our understanding of social phenomena. Part IV, “Power, dependence and social interaction”, concentrates on power and dependence as the essential building blocks of social interaction – so much so, that they are found to be relevant for the most disparate domains and levels of analysis, from formal models to socio-cognitive theories, via social simulations. Vittorio Pelligra (chapter 17) identifies the greatest limit of game theory in its poor understanding of intersubjectivity, reviews empirical data highlighting the relevance of mentalizing and empathy in strategic interaction, and argues that conceptualizing a hierarchy of higher order beliefs in psychological game theory improves our formal understanding of the motivational structure of real social agents. Davide Grossi and Paolo Turrini (chapter 18) build on Castelfranchi’s work on dependence theory to develop a full-fledged game-theoretical analysis of social interaction in terms of dependence and related notions, such as dependence cycle and reciprocity, thus revealing a close, unexpected kinship between game theory and dependence theory. Helder Coelho (chapter 19) puts the notion of power, in particular social power, in contact with that of leadership, distinguishes power-over from power-of, and dwells on the implications of such an analysis for understanding and transforming contemporary societies, be they real or artificial. Raffaella Pocobello and Tarek el Sehity (chapter 20) illustrate how Castelfranchi's goal-centered approach to cognition sheds light on the mental components of a new paradigm in the field of mental health, recovery, understood as the individual's capacity to develop a meaningful life and a self-concept beyond the illness, with a strong emphasis on autonomy and empowerment.

Part V, "Trust & delegation", deals with the theory of trust and delegation, one of the most influential contributions of Castelfranchi and his team to multi-agent systems and social science. Rino Falcone and Maria Miceli (chapter 21) analyze the complex relationships between trusting and being trustworthy, with special emphasis on how the fact that agent X trusts agent Y might be perceived by Y as a sign of X's trustworthiness, and how being trusted will also increase the likelihood that Y proves in fact to be worthy of that trust. Andreas Herzig, Emiliano Lorini and Frédéric Moisan (chapter 22) propose a simple logic of belief and action that allows to express the concepts of belief, goal, ability, willingness, and opportunity, upon which the socio-cognitive theory of trust is built, and then provide a decision procedure and a proof of completeness for such logic. Serena Villata, Guido Boella, Dov Gabbay and Leendert van der Torre (chapter 23) develop a cognitive model of conflicts in trust using argumentation, in which trust serves to minimize the uncertainty in the interactions of information sources, while argumentation is used to reason about trust and its two main dimensions, competence and sincerity. Elisabetta Erriquez, Wiebe van der Hoek and Michael Wooldridge (chapter 24) take distrust, rather than trust, as their primitive concept, and use it to model how agents in a society may form stable coalitions based on their mutually perceived level of trustworthiness, or lack thereof – an approach that they successfully apply to a fascinating case of complex trust and distrust relationships, namely, Shakespeare's *Othello*. Patrick Doherty and John-Jules Meyer (chapter 25) focus on the logic of delegation, proposing to extend its application, typically limited to multi-agent systems and social sciences, to collaborative robotic systems, by instantiating delegation as a speech act and then illustrating its usefulness in a running prototype, used in collaborative missions with multiple unmanned aerial vehicle systems.

Part VI, "Communication", brings together many different strands of Castelfranchi's work, such as persuasion, deception, gestures, and behavioral implicit communication, to highlight their mutual connections and shared roots. Oliviero Stock and Marco Guerini (chapter 26) take on Castelfranchi's frequent invitation to understand the ethical implications of building intelligent machines, *before* we build them, and apply it to persuasive systems, offering both a bird's eye view and a critical assessment of this thriving research area. Sebastiano Bagnara and Simone Pozzi (chapter 27) discuss the import of Castelfranchi's notion of behavioral implicit communication for the design of Natural User Interfaces, arguing that traces of human activity could be used to teach such interfaces what are the relevant gestures and what is their intended meaning for users – two aspects on which current systems are sadly lacking. Michele Piunti, Alessandro Ricci and Luca Tummolini (chapter 28) introduce a vision of agent-oriented Ambient Intelligence (AmI) systems, understood as not only mirroring but also augmenting the physical world, and discuss how to enable such systems to detect and digitally represent the traces left by humans in the physical world, in order to fully exploit the value of stigmergy as a coordination mechanism. The relevance of stigmergy for intelligent coordination is also the focus of Andrea Omicini (chapter 29), who reviews and classifies some of the main types of environment-based coordination, to discuss their impact on the engineering of self-organising socio-technical systems. Swati Gupta, Kayo Sakamoto and Andrew Ortony (chapter 30) endeavor to provide a comprehensive and systematic account of the ubiquitous phenomenon of verbal deception, building on the work done by Castelfranchi and many others on this topic, and ending up with two original taxonomies, one for the types of verbal deception, and one for the strategies used to verbally deceive.

Part VII, "Norms, organizations and institutions", focuses on another central aspect of Castelfranchi's theorizing: the nature, dynamics and evolution of norms and normative reasoning, as well as their role in the emergence of institutions and in the functioning of organizations. Pivotal to that theory is the notion of commitment, which is the topic addressed by Munindar Singh (chapter 31): he first provides a comprehensive and authoritative survey of the rich literature on this topic, with special focus on commitments in multi-agent systems, highlighting key concepts, lingering confusions, and promising directions for future research; then he endeavors to present the main points of disagreements between his own views and those of Castelfranchi, in spite of their largely shared background and interests. Giovanni Sartor (chapter 32) discusses norm compliance, arguing that complex normative systems, albeit very successful at directing people's thoughts and actions, cannot be, as a whole, objects of the individual agents' mental attitudes. Still, for Sartor this feature is not evidence against the mentalistic theory of norms developed by Castelfranchi and collaborators, if one acknowledges that the agents adopt a general policy-based intention to comply with the normative system as a whole, which can be based on different mental attitudes, ranging from self-interest to pro-social motivations. Giulia Andrighetto, Rosaria Conte and Francesca Giardini (chapter 33) use a simulation-based methodology to model how cognitive activities and representations affect institutional change, and play a decisive role in their selection and retention: as a relevant case study, they focus on enforcing institutions that are hypothesized to have evolved from retaliatory to punishing, and even sanctioning, systems, thanks to and by means of specific cognitive capacities. Jaime Simão Sichman (chapter 34) articulates and defends the idea that autonomous cognitive agents, immersed in an open environment, are more efficient and adaptive to changes if they can represent, elaborate and exploit information about other agents and organizations, since this enables a virtuous loop between agents interactions, coalitions, and organizations. Frank Dignum and Virgina Dignum (chapter 35) investigate how norms about having emotions, as well as sanctions for their violation, can exist and make sense, in spite of the fact that emotions are something that people cannot (easily) control, and therefore a very strange object of normative concern: nonetheless, they provide a formal description of these kinds of norms, to discuss whether they are the same as other norms or have special properties.

Part VIII, “Cognitive and computational social science”, takes a more methodological and meta-theoretical stance on Castelfranchi’s work, to consider the implications of agent-based modeling and simulation for the understanding of cognitive and social phenomena. Domenico Parisi (chapter 36) argues that scientific theories should be formulated and presented as computer-based artifacts, rather than verbally or even mathematically, in order to remove ambiguity, overcome disciplinary fragmentation, avoid value-laden implications and biases, and facilitate technological transfer and socially relevant application of those theories: he then exemplifies the benefits and potentialities of this method with his ongoing work on evolutionary robotics, discussing what we can learn about motivations, emotions, mental life, language, social interaction, economic dynamics, politics, and even culture, by trying to evolve robots capable of manifesting such phenomena in their artificial ecology. Mario Paolucci (chapter 37) presents social simulation as one of our best chances for breakthroughs in understanding society, and yet also discusses why social simulation so far largely failed to deliver substantial results on this big challenge: then he introduces the concept of crowdsourcing, elaborating on how it could positively reshape this methodology for computational social science. Luis Antunes (chapter 38) describes the epistemological challenges and risks entailed by a simulation-based research program, with an emphasis on what types of mistakes are most likely to occur at various stages of the process (from conceptual analysis to modeling, from simulation design to actual implementation, and on), and what methods are available to predict, minimize and correct them – an enterprise that, according to Antunes, has been central to Castelfranchi’s approach, often focused on foundational issues both in multi-agent systems and in social science. Finally, Yurij Castelfranchi (chapter 39), starting from a small digression on the infamous micro-macro, agency-structure dilemma of social sciences, shows how insights coming from multi-agent systems and the theory of social functions could improve our understanding of two key societal issues: the politics and economy of science in the contemporary regime of knowledge production, and the functions and issues of the public communication of science and technology. This also uncovers a common core between Cristiano Castelfranchi’s theoretical contributions (in particular, his ideas on social emergence and “immergence”) and his political positions on science communication and science policies[4].

This last contribution is especially tinged with affection, which is no surprise, since the author is Castelfranchi's eldest son, a researcher himself (in sociology), as well as a science writer, journalist, traveler, and more – in sum, a true Castelfrankian. Exactly like his younger brother, Vania Castelfranchi, also a great traveler, as well as a professional actor, director, author, mime, juggler, teacher, etc. It is thanks to Vania that this volume is graced by such a nice cover, which we would have never had the talent to design or the competence to realize. Finally, our little conspiracy in putting together this volume, unbeknownst to Cristiano, would have not been possible without the generous help of Rosanna Bosi, psychotherapist, pedagogist, instrumentalist, and, most importantly, the love of his life. If Cristiano had edited this volume himself, he would have certainly dedicated it to her, as he did so often in the past. This time, we will take the liberty of acting on his behalf, and dedicate this volume for Cristiano to Ros, the worthy companion of our wonderful teacher, inspired colleague, and dear friend.

Rome, November 2012

THE EDITORS

[4] As mentioned, the volume is completed by Cristiano Castelfranchi's own contribution, in Part IX, "Afterword": there he offers an impressive summary of almost 40 years of research on goals and goal-directed behavior. However, we briefly sketched the contents of that chapter at the very beginning of this Introduction, and we would never presume to summarize his monumental contribution in just a few lines, nor we dare trying. Thus, no more is said about it here, except for a strong suggestion to readers to peruse Cristiano's paper *before* anything else, especially if they are not yet familiar with his work. Doing so, in fact, will offer them a favored standpoint to appreciate all other contributions, as well as a fresh angle on cognition in general.