Introduction

The aim of this book, carried out in quite a user-friendly way, is clear from its title: to systematize a general *theory* of 'trust'; to provide an organic *model* of this very complex and dynamic phenomenon on cognitive, affective, social (interactive and collective) levels.

Why approach such a scientific project, not only from the point of view of Cognitive and Behavioral Sciences, but also from Artificial Intelligence (AI) and in particular 'Agent' theory domains? Actually, trust for Information and Communication Technologies (ICT) is for us just an application, a technological domain. In particular, we have been working (with many other scholars)[1] in promoting and developing a tradition of studies about trust with Autonomous Agents and in Multi-Agent Systems (MAS). The reason is that we believe that an AI oriented approach can provide – without reductionisms – good systematic and operational instruments for the explicit and well-defined representation of goals, beliefs, complex mental states (like expectations), and their dynamics, and also for modeling social action, mind, interaction, and networks. An AI approach with its programmatic 'naiveté' (but being careful to avoid simplistic assumptions and reductions of trust to technical tricks – see Chapter 12) is also useful for revising the biasing and distorting 'traditions' that we find in specific literature (philosophy, psychology, sociology, economics, etc.), which is one of the causes of the recognized 'babel' of trust notions and definitions (see below, Section 0.2).

However, our 'tradition' of research at ISTC-CNR (Castelfranchi, Falcone, Conte, Lorini, Miceli, Paglieri, Paolucci, Pezzulo, Tummolini, and many collaborators like Poggi, De Rosis, Giardini, Piunti, Marzo, Calvi, Ulivieri, and several others) is a broader and Cognitive

[1] We are grateful to our collegues and friends in the AI Agent community discussing these issues with us for the last 10 years: Munindar Singh, Yao-Hua Tan, Suzanne Barber, Jordi Sabater, Olivier Boissier, Robert Demolombe, Andreas Herzig, Andrew Jones, Catholijn Jonker, Audun Josang, Stephen Marsh, Carles Sierra. And also to other colleagues from different communities, like Michael Bacharach, Sandro Castaldo, Michele Costabile, Roderick Kramer, Vittorio Pelligra, Raimo and May Tuomela. The following articles have been reproduced in this book: Cristiano Castelfranchi and Rino Falcone, *Principles of trust for MAS: Cognitive Anatomy, Social Importance, and Quantification,* Proceedings of the International Conference on Multi-Agent Systems (ICMAS'98), Paris, July, pp.72–79 (1998). Reproduced by Permission of ©1998 IEEE. Cristiano Castelfranchi and Rino Falcone, *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, 31(5): 406–418, September 2001. Reproduced by Permission of ©2001 IEEE

Trust Theory: A Socio-Cognitive and Computational Model Cristiano Castelfranchi and Rino Falcone
 ISBN: 978-0-470-02875-9

Science-oriented tradition: to systematically study the *'cognitive mediators of social action'*: that is, the mental representations supporting social behaviors and collective and institutional phenomena; like: cooperation, social functions, norms, power, social emotions (admiration, envy, pity, shame, guilt, etc.).

Thus, trust was an unavoidable and perfect subject: on the one hand, it is absolutely crucial for social interaction and for collective and institutional phenomena (and one should explain 'why'); on the other hand, it is a perfect example of a necessary cognitive 'mediator' of sociality, and of integration of mind and interaction, of epistemic and motivational representations, of reasoning and affects. Our effort is in this tradition and frame (see below, Section 0.3).

Respecting and Analyzing Concepts

Quite frequently in science (especially in the behavioral and social sciences, which are still in search of their paradigmatic status and recognition) '*Assimilation*' (in Piaget's terms)[2] prevails on '*Accommodation*'.

That is, the *simplification* of factual data, the *reduction* of real phenomena in order they fit within the previously defined 'schemes', and in order to confirm the existing theories with their conceptual apparatus, strongly prevails on the adjustment of the concepts and schemes to the complexity and richness of the phenomenon in object.

In such a way, well-defined (and possibly formalized) schemes become blinkers, a too rigid and arbitrary filter of reality. Paradoxically reality must conform to theory, which becomes not just – as needed – abstract, parsimonious, 'ideal-type', and 'normative', but becomes 'prescriptive'. Scholars no longer try to develop a good general theory of 'trust' as conceived, used, perceived in 'natural' (cultural) contexts; they prescribe what 'trust' *should* be, in order to fit with their intangible theoretical apparatuses and previous defined basic notions. They deform their object by (i) pruning what is not interesting for their discipline (in its consolidated current asset), and by (ii) forcing the rest in its categories.

In this book, we try to assume an 'Accommodation' attitude.[3] For three reasons:

First, because the current trust 'ontology' is really a recognized mess, not only with a lot of domain-specific definitions and models, but with a lot of strongly contradictory notions and claims.

Second, because the separation from the current (and useful) notion of trust (in common sense and languages) is too strong, and loses too many interesting aspects and properties of the social/psychological phenomenon.

Third, because we try to show that all those ill-treated aspects not only deserve some attention, but are much more coherent than supposed, and can be unified and grounded in a principled way.

[2] See, for simplicity: http://www.learningandteaching.info/learning/assimacc.htm; http://projects.coe.uga.edu/epltt/index.php?title=Piaget%27s_Constructivism.

[3] Notice that both attitudes are absolutely natural and necessary for good cognitive development and adaptation; stabilizing categories and schemes; adjusting them when they become too deforming or selective.

The Characteristics of Our Trust Model

What we propose (or try to develop) is in fact a model with the following main features:

1) An *integrated model*
   - A definition/concept and a pre-formal 'model' that is composite or better layered: with various constituents and a core or kernel. A model that is able to assemble in a principled way 'parts' that are usually separated or lost for mere disciplinary and reductive interests.
   - Not a summation of features and aspects, but a 'gestalt'; a complex *structure* with specific components, relations, and functions.
   - A model apt to explain and justify in a coherent and non *ad hoc* way the various properties, roles, functions, and definitions of 'trust'.
2) *A socio-cognitive model*
   Where 'cognitive' does not mean 'epistemic' (knowledge), but means 'mental' (explicit mental representations); including motivational representations (various goal families).
   Trust should not be reduced to epistemic representations, 'beliefs' (like in many definitions that we will discuss: grounded prevision; subjective probability of the event; strength of the belief; statistic datum; etc.). Our 'integration' is architectural and 'pragmatic' where beliefs are integrated with *motivation* (goals and resultant affectivity, which is goal-based) and with *action*, and the consequential social effects and relations.
3) An *analytic* and *explicit* model
   Where the various components or 'ingredients' (epistemic, motivational, of action, and relational) are represented in an explicit format, ready to be formalized, and so on. And based on a 'normative' ('ideal-typical') frame in terms of those explicit mental constituents. However, there is a clear claim that this is just the prototypical model, the 'ideal' reference for analytical reasons. But *there are implicit and basic forms of trust*, either routine-based, mindless, and automated, or merely 'felt' and affect-based forms. In these 'implicit' forms, the same 'constituents' are just potentially present or are present in a tacit, procedural way; just primitive forerunners of the explicit advanced representations, but with the same functions: equifinal. This is, for example, the distinction between the true 'cognitive evaluation' and the 'affective appraisal' (see Chapter 8).
4) A *multi-factor and multi-dimensional* model of trustworthiness and of trust, and a *recursive* one.
   Where trust in agent Y is based on beliefs about its powers, qualities, capacities; which actually are the basis for the global trust in Y, but also are sub-forms of trust: trust in specific virtues of Y (like 'persistence', 'loyalty', 'expertise', etc.).
5) A *dynamic* model
   Where trust is not just a fixed attitude, or a context independent disposition, or the stable result of our beliefs and expectations about Y. But is context dependent, reasoning dependent, self-feed; and also reactive and interactive. There are two kinds of dynamics: one is 'internal' (mind-decision-action); the other is 'external': the dynamics of interactive, relational, and network trust links. And, not forgetting, they are intertwined.
6) A *structurally related notion*
   On such a basis, one should provide an explicit, justified, and systematic theory of the relationships between the notion/phenomenon of trust and other strongly related notions/phenomena: previsions, expectations, positive evaluations, trustworthiness,

uncertainty and risk, reliance, delegation, regularities and norms, cooperation, reputation, safety and security, and so on; and correlated emotions: relaxation and feeling safe, surprise, disappointment, betrayal, and so on.

7) *A non-prescriptive* model
We do not want to claim *'This is the "real", right, meaning; the correct use. The other current uses are mistaken or inappropriate uses of common languages'*. For example: *'Trust is just based on regularities and rules; even when the prevision is something that I do not like/want, that I worry about', or: 'The only true trust is the moral and personal one; the one that can be betrayed';* or again: *'There is no trust when there are contracts, laws, authorities involved'; 'Trust is there only in reciprocal and symmetric situations';* and so on.

Our aim is not to abuse the concept, but at the same time to be able to situate in a precise and justified way a given special condition or property (like: 'grounded prediction') as a possible (and frequent) sub-component and usual basis of trust; or to categorize the moral-trust or the purely personal trust as (important) *types* of interpersonal trust.

The Structure of the Book

We start with a 'landscape' of the definitional debate and confusion; and with the discussion of some important definitions containing crucial ingredients. We try to show how and why some unification and abstraction is possible.

In Chapter 2 we present in a systematic way our basic model. How trust is not only a disposition or a set of beliefs (evaluation or prediction), but also a decision to rely on, and the following 'act' of, and the consequential social relation; and how these layers are embedded one into the other. How trust is not only about 'reliability' but also about 'competence', and about feeling safe, not being exposed to harms. How trust presupposes specific mental representations: evaluations, expectations, goals, beliefs of 'dependence', etc. How trust implies an 'internal' attribution to the trustee, based on external cues. How there are broader notions and more strict (but coherent) ones: like 'genuine' trust, relying on the other's goal-adoption (help), or trust relying on his 'morality'.

In Chapter 3 we present the quantification of trust. How trust has various strengths and degrees (precisely on the basis of its constituents: beliefs, goals). How trust enters the decision to delegate or not to delegate a task. How it copes with perceived risk and uncertainty. How we can say that trust is too great or too little.

In Chapter 4 we try to better understand the trust concept analyzing strictly related notions like: lack of trust, mistrust, diffidence. We also consider and develop the role of implicit trust: so relevant in many social actions.

In Chapter 5 we consider the affective trust. Even if in this book the emotional trust is (deliberately) a bit neglected, we briefly analyze this aspect and evaluate its relevance and show its interactions and influences with the more rational (reason-based) part.

In Chapter 6 trust dynamics is presented in its different aspects: how trust changes on the basis of the trustor's experiences; how trust is influenced by trust; how diffuse trust diffuses trust; how trust can change using generalization reasoning.

In Chapter 7 we consider the very interesting relationships between trust, control and autonomy, also with respect to the potential autonomy adjustments. In particular we show how very often, in the relationships between trust and control, some relevant aspects are neglected.

In Chapter 8 we present our deep disagreement with that economical point of view which reduces trust to a trivial quantification and measure of the costs/benefits ratio and risks, sacrificing a large part of the psychological and social aspects.

In Chapter 9 we underline the role of trust in Social Order, both as institutional, systemic glue producing shared rules, and as spontaneous, informal social relationships. In fact, we present trust as the basis of sociality.

In Chapter 10 we change the point of view in the trust relationship, moving to the trustee's side and analyzing how its own trustworthiness can be exploited as a relational capital. We consider in general terms the differences between simple dependence networks and trust networks.

In Chapter 11 we show a fuzzy implementation of our socio-cognitive model of trust. Although very simple and reduced, the results of the implementations present an interesting picture of the trust phenomenon and the relevance of a socio-cognitive analysis of it.

In Chapter 12 we present the main technological approaches to trust with their merits and limits. The growth of studies, models, experiments, research groups, and applications show how much relevance trust is gaining in this domain. How the bottleneck of technology can be measured by the capacity of integrating effective social mediators in it.

In Chapter 13 we draw conclusions and also present a potential challenge field: the interactions between neuro-trust (referring to the studies on the neurobiological evidence of trust) and the theoretical (socio-cognitive) model of trust: without this interaction the description of the phenomenon is quite poor, incomplete and with no prediction power.

For a schematic view of the main terms introduced and analyzed in this book see the Trust, Theory and Technology site at http://www.istc.cnr.it/T3/.