TRUST THEORY

A SOCIO-COGNITIVE AND COMPUTATIONAL MODEL

Cristiano Castelfranchi

Institute of Cognitive Sciences and Technologies (ISTC) of the Italian National Research Council (CNR), Italy

Rino Falcone

Institute of Cognitive Sciences and Technologies (ISTC) of the Italian National Research Council (CNR), Italy

WILEY

A John Wiley and Sons, Ltd., Publication

TRUST THEORY

Wiley Series in Agent Technology

Series Editor: Michael Wooldridge, *University of Liverpool, UK*

The 'Wiley Series in Agent Technology' is a series of comprehensive practical guides and cutting-edge research titles on new developments in agent technologies. The series focuses on all aspects of developing agent-based applications, drawing from the Internet, telecommunications, and Artificial Intelligence communities with a strong applications/technologies focus.

The books will provide timely, accurate and reliable information about the state of the art to researchers and developers in the Telecommunications and Computing sectors.

Titles in the series:

Padgham/Winikoff: Developing Intelligent Agent Systems 0-470-86120-7 (June 2004)

Bellifemine/Caire/Greenwood: Developing Multi-Agent Systems with JADE 0-470-05747-5 (February 2007)

Bordini/Hübner/Wooldrige: Programming Multi-Agent Systems in AgentSpeak using Jason 0-470-02900-5 (October 2007)

Nishida: Conversational Informatics: An Engineering Approach 0-470-02699-5 (November 2007)

Jokinen: Constructive Dialogue Modelling: Speech Interaction and Rational Agents 0-470-06026-3 (April 2009)

TRUST THEORY

A SOCIO-COGNITIVE AND COMPUTATIONAL MODEL

Cristiano Castelfranchi

Institute of Cognitive Sciences and Technologies (ISTC) of the Italian National Research Council (CNR), Italy

Rino Falcone

Institute of Cognitive Sciences and Technologies (ISTC) of the Italian National Research Council (CNR), Italy

WILEY

A John Wiley and Sons, Ltd., Publication

This edition first published 2010

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Castelfranchi, Cristiano.
Trust theory : a socio-cognitive and computational model / Cristiano Castelfranchi, Rino Falcone.
p. cm.
Includes index.
ISBN 978-0-470-02875-9 (cloth)
1. Trust. 2. Trust–Simulation methods. 3. Artificial intelligence–Psychological aspects. 4. Cognitive science.
I. Falcone, Rino. II. Title.
BF575.T7C37 2010
302'.1–dc22

2009040166

A catalogue record for this book is available from the British Library.

ISBN 9780470028759 (H/B)

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India
Printed and Bound in Singapore by Markono.

This book is dedicated
to Mario, Ketty, Eugenio, and Maria, our roots,
to Rosanna and Ivana, our life companions;
to Yurij, Vania and Giulio, our realized dreams;
to the many collegues who consciously or unconsciuously contributed to the ideas included in it;
to our Country, better to the Idea of Country, of Collective, of Institutional Entity, in which as such it is possible develop socially centred hopes, ambitions, dreams, so contributing to the dignity and future for any individual.

Contents

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/.

Foreword

I turn up to give a lecture at 9 am on a Monday morning, trusting that my students will attend; and they in turn reluctantly drag themselves out of bed to attend, trusting that I will be there to give the lecture. When my wife tells me that she will collect our children from school, I expect to see the children at home that night safe and sound. Every month, I spend money, trusting that, on the last Thursday of the month, my employer will deposit my salary in my bank account; and I trust my bank to safeguard this money, investing my savings prudently. Sometimes, of course, my trust is misplaced. Students don't turn up to lectures; my bank makes loans to people who have no chance of repaying them, and as a consequence they go bankrupt, taking my savings with them. But despite such disappointments, our lives revolve around trust: we could hardly imagine society functioning without it.

The rise of autonomous, computer-based agents as a technology gives trust an interesting new dimension. Of course, one issue is that we may not be comfortable trusting a computer program to handle our precious savings. But when software agents interact with people an entirely new concern arises: why or how should a computer program trust 'us'? How can we design computer programs that are safe from exploitation by un-trustworthy people? How can we design software agents that can understand how trust works in human societies, and live up to human expectations of trust? And what kind of models of trust make sense when software agents interact with 'other' software agents?

These considerations have led to attempts by cognitive scientists, computer scientists, psychologists, and others, to develop models of trust, and to implement these tentative models of trust in computer programs. The present book is the first comprehensive overview of the nascent field of modeling trust and computational models of trust. It discusses trust and the allied concept of reputation from a range of different backgrounds. It will be essential reading for anybody who wants to understand the issues associated with building computer systems that work with people in sensitive situations, and in particular for researchers in multi-agent systems, who will deploy and build on the techniques and concepts presented herein. The journey to understand trust from a scientific, technological, and computational perspective may only just have begun, but this book represents a critical milestone on that journey.

Michael Wooldridge

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

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/.

1

Definitions of Trust: From Conceptual Components to the General Core

In this chapter we will present a thorough review of the predominant definitions of trust in the literature, with the purpose of showing that, in cognitive and social sciences, there is not yet a shared or prevailing, and clear and convincing notion of *trust*. Not surprisingly, this appalling situation has engendered frequent and diffuse complaints.[1] However, the fact that the use of the term *trust* and its analytical definition are confused and often inaccurate should not become an unconscious alibi, a justification for abusing this notion, applying it in any *ad hoc* way, without trying to understand if, beyond the various specific uses and limited definitions, *there is some common deep meaning, a conceptual core to be enlightened*.

On the contrary, most authors working on trust provide their own definition, which frequently is not really general but rather tailored for a specific domain (commerce, politics, technology, organization, security, etc.). Moreover, even definitions aimed at being general and endowed with some cross-domain validity are usually incomplete or redundant: either they miss or leave implicit and give for presupposed some important components of trust, or they attribute to the general notion something that is just accidental and domain-specific.

The consequence is that there is very little overlapping among the numerous definitions of trust, while a strong common conceptual kernel for characterizing the general notion has yet to emerge. So far the literature offers only partial convergences and 'family resemblances' among different definitions, i.e. some features and terms may be common to a subset of definitions but not to other subsets.

This book aims to counteract such a pernicious tendency, and tries to provide *a general, abstract, and domain-independent notion and model of trust*.

[1] See for example Mutti (1987: 224): 'the number of meanings attributed to the idea of trust in social analysis is disconcerting. Certainly this deplorable state of things is the product of a general *theoretical negligence*. It is almost as if, due to some strange self-reflecting mechanism, social science has ended up losing its own trust in the possibility of considering trust in a significant way'.

Trust Theory Cristiano Castelfranchi and Rino Falcone

This theoretical framework:

- should take inspiration from and further analyze the common-sense notion of trust (as captured by natural languages), as well as the intuitive notions frequently used in the social sciences, but
- should also define a technical scientific construct for a precise characterization of trust in cognitive and social theory, while at the same time
- accounting for precise relationships with the most important current definitions of trust, in order to show what they all have in common, regardless of their different terminological formulations.

We believe this generalization and systematization to be both possible and necessary. In this chapter, we will start identifying the most recurrent and important features in trust definitions, to describe them and explain their hidden connections and gaps. This will be instrumental to a twofold purpose: on the one hand, we will show how our layered definition and quite sophisticated model can account for those features of trust that appear to be most fundamental; on the other hand, we will discuss why other aspects of current definitions of trust are just local, i.e. relevant only for a very specific problem or within a restricted domain. In this analysis, we will take as initial inspiration Castaldo's content analysis of trust definitions (Castaldo, 2002).

This critical effort will serve both to clarify the distinctive features of our own perspective on trust, and to highlight the most serious limitations of dominant current approaches.

1.1 A Content Analysis

In dealing with the current 'theoretical negligence' and conceptual confusion in trust definitions, Castaldo (Castaldo, 2002) applied a more descriptive and empirical approach, rather different but partially complementary to our own. Castaldo performed a content analysis of 72 definitions of trust (818 terms; 273 different terms), as employed in the following domains: Management (46%), Marketing (24%), Psychology (18%), and Sociology (12%). The survey covered the period from the 1960s to the 1990s, as described in Table 1.1:

Table 1.1 Number of trust definitions in different periods

Year	Definitions	Fraction
1960–69	4	(5.6%)
1970–79	5	(7.0%)
1980–89	19	(26.4%)
1990–99	44	(51.0%)
Total	72	(100.0%)

This table is from Castaldo. For more sophisticated data and comments, based on cluster analysis, see (Castaldo, 2002).

Source: Reproduced with kind permission of © 2002 Società editrice il Mulino.

This analysis is indeed quite useful, since it immediately reveals the degree of confusion and ambiguity that plagues current definitions of trust. Moreover, it also provides a concrete framework to identify empirically different 'families' of definitions, important conceptual nucleuses, necessary components, and recurring terms. Thus we will use these precious results as a first basis for comparison and a source of inspiration, and only later will we discuss in detail specific definitions and models of trust.

Castaldo summarizes the results of his analysis underlining how the trust definitions are based on five inter-related categories. They are:

- The *construct*, where trust is conceived 'as an *expectation*, a *belief*, *willingness*, and an *attitude*' (Castaldo, 2002).
- The *trustee*, 'usually individuals, groups, firms, organizations, sellers, and so on' (Castaldo, 2002). Given the different nature of the trustee (individuals, organizations, and social institutions), there are different types of trust (personal, inter-organizational and institutional). These trustees 'are often described by reference to different characteristics in the definitions being analyzed – specific competencies, capacities, non-opportunistic motivations, personal values, the propensity to trust others, and so on' (Castaldo, 2002).
- *Actions* and *behaviors*, as underlined also from other authors (e.g. (Moorman Zaltman and Desphande, 1992)) the behavioral aspect of trust is fundamental for 'recognizing the concept of trust itself' (Castaldo, 2002); both trustor and trustee behaviors have to take into account the consistence of the trust relationship. Behavioral aspects of trust have been studied also showing its multi-dimensional nature (e.g (Cummings and Bromiley, 1996)).
- *Results* and *outputs* of behavior, trustee's actions are presumed to be both predictable and positive for the trustor. 'The predictability of the other person's behavior and the fact that the behavior produces outcomes that are favorable to the trustor's objectives, are two typical results of trust. This has been particularly studied in works which suggest models designed to identify the consequences of trust (e.g. (Busacca and Castaldo, 2002)) (Castaldo, 2002).
- The *risk*, without uncertainty and risk there is no trust. The trustor has to believe this. They have to willingly put themselves into a 'position of vulnerability with regard to the trustee'. Risk, uncertainty and ambiguity (e.g. (Johannisson, 2001)) are the fundamental analytic presuppositions of trust, or rather the elements that describe the situations where trust has some importance for predictive purposes. (. . .).

[There is some sort of] logical sequence (. . .) [which has] often been suggested in the definitions. This sequence often regards trust as the *expectation, belief* (and so on) that a subject with specific characteristics (honesty, benevolence, competencies, and so on) *will perform* actions designed to produce *positive results* in the future for the trustor, in situations of consistent *perceived risk* (Castaldo, 2002).

Notwithstanding its merits, the main limit of Castaldo's analysis is that it fails to provide a stronger account of the *relationships* among these recurrent terms in trust definitions, i.e. indicating when they are partial synonyms, rather than necessary interdependent parts of a larger notion, or consequences of each other, and so on. Just an empirical, descriptive and co-relational account remains highly unsatisfactory. For example, it is true that 'Trust has been predominantly conceived as an *expectation*, a *belief*, *willingness*, and an *attitude*'.

However, it remains to be understood what are the conceptual ties between *belief* and *expectation*, or between *belief* and *willingness* (is one a species of the other? Does one

concept contain the other?). What are their exact roles in the processing of trust: For instance, what is the procedural relationship (e.g. sequential) between *belief* and *willingness*, which certainly is not a kind of belief? And why do some authors define trust *only* as a *belief*, while other authors only consider it as *willingness* and as a *decision* or *action*? Statistical relations do not even begin to address these questions.

An in-depth analysis of the *conceptual interconnections* among different facets of trust is also instrumental to achieve *a more adequate characterization of this notion*, since a good definition should be able to cover these different aspects and account for their relevance and their mutual relationships, or motivate their exclusion.

In particular, any theoretical model should take into account that trust is a *relational* construct, involving at the same time:

- A subject ***X*** (*the trustor*) which necessarily is an 'intentional entity', i.e. a system that we interpret according to Dennett's intentional stance (Dennett, 1989), and that is thus considered a cognitive agent.
- An addressee ***Y*** (*the trustee*) that is an *agent* in the broader sense of this term (Castelfranchi, 1998), i.e. an entity capable of causing some effect as the outcome of its behavior.
- The causal process itself (*the act,* or *performance*) and its result; that is, an act α of ***Y*** possibly producing the desired outcome ***O***.

Moreover, we should also never forget that trust is a *layered notion*, used to refer to several different (although interrelated) meanings (see Chapter 2):

- in its basic sense, trust is just a mental and affective *attitude* or *disposition* towards ***Y***, involving two basic types of *beliefs*: *evaluations* and *expectations*;
- in its richer use, trust is a *decision* and *intention* based on that disposition;
- as well as the *act* of *relying* upon ***Y***'s expected behavior;
- and the consequent social *relation* established between ***X*** and ***Y***.

If we now apply this analysis to the results summarized in Table 1.2, we can make the following observations:

- As for the terms ***Will***, ***Expect***, ***Belief***, ***Outcome***, ***Attitude***, they match the relation we postulate quite closely: *will* refers to the future (as Castaldo emphasizes), thus it is also included in the notion of *expectation*, which in turn involves a specific kind of *belief*: in its minimal sense, an expectation is indeed a belief about the future (Miceli and Castelfranchi, 2002; Castelfranchi and Lorini, 2003; Castelfranchi, 2003). Moreover, the term *belief* implies a mental attitude, and we can say that trust as evaluation and expectation is an *attitude* towards the trustee and his action: the *outcome*, the events, the situation, the environment.
- As for the terms ***Action*** and ***Decision***, they refer to trust as the deciding process of *X* and the subsequent *Y*'s course of action; hence they are general, but only with reference to the second and richer meaning of trust discussed above (see also below and Chapter 2).
- As for the terms ***Expect***, ***Outcome***, ***Rely***, ***Positive***, ***Exploit***, and ***Fulfill***, again they are tightly intertwined according to our relational view of trust: the positive outcome of the trustee's action is expected, relied upon, and exploited to fulfill the trustor's objective. In short: *X has*

Table 1.2 Most frequently used terms in trust definitions[2]

Terms	Frequency
Subject (Actor, Agent, Another, Company, Customer, Firm, Group, Individual, It, One, Other, Party, People, Person, Salesperson, Somebody, Trustee, Trustor)	180
Action (Action, Act, Behavior, Behave, Behavioral)	42
Will	29
Expect, Expectation, Expected, Expectancy	24
Belief, Believe	23
Outcome, Result, Performance, Perform	19
Rely, Reliable, Reliance, Reied, Reliability, Relying	18
Trust, Trusting, Trustworthy	17
Confident, Confidence	16
Willingness, Willing	14
Take, Taken, Taking, Accept, Accepted, Acceptable	11
Risk, Risky, Risking	11
Vulnerable, Vulnerability	11
Relationship	10
Exchange	9
Based	8
Competent, Competence, Capabilities	7
Positive	7
Cooperate, Cooperation, Coordination	6
Exploit, Exploitation	6
Situation	6
Attitude	5
Decide, Decision	5
Fulfill, Fulfilled, Fulfillment	5
Held	5
Intention, Intentionally, Intend	5
Involve, Involved, Involvement, Involving	5
Mutual, Mutually	5
Word	5
Would	5

Source: Reproduced with kind permission of © 2002 Società editrice il Mulino.

a goal (a desire or need) that is expected to be fulfilled thanks to Y's act; X intends to exploit the positive outcome of Y's act, and relies upon Y for fulfilling the goal.

- As for the terms ***Taken***, ***Accept***, ***Risk***, and ***Vulnerable***, their relationship is that while deciding to count on *Y*, to trust *Y* (according to trust as decision), *X* is necessarily accepting the risk of becoming *vulnerable* by *Y*, since there is uncertainty both in *X*'s knowledge (incomplete, wrong, static) and in the (unpredictable, unknown) dynamics of the world.

[2]This table is from Castaldo. For more sophisticated data and comments, based on cluster analysis, see (Castaldo, 2002).

Whenever deciding to depend on Y for achieving O, X is exposed both to failure (not fulfilling O) and to additional harms, since there are intrinsic costs in the act of reliance, as well as retreats to possible alternatives, potential damages inflicted by Y while X is not defended, and so on. As we will discuss more thoroughly in the next chapters, all these risks are direct consequences of X's decision to trust Y.

- As for the terms ***Competence*** and ***Willingness***, they identify the two basic prototypical features of 'active'[3] *trust in Y*, i.e. the two necessary components of the positive evaluation of Y that qualify trust:

  - The belief of X (evaluation and expectation) that Y is *competent* (able, informed, expert, skilled) for effectively doing α and produce O;
  - The belief of X (evaluation and expectation) that Y is *willing* to do α, intends and is committed to do α – and notice that this is precisely what makes an agent Y predictable and reliable for X. Obviously this feature holds only when Y is a cognitive, intentional agent. It is in fact just a specification of a more abstract component that is Y's *predictability*: the belief that 'Y will actually do α and/or produce O', contrasted with merely having the potentiality for doing so.

In sum, a good definition of trust, and the related analytical model that supports it, must be able to explicitly account for two kinds of relationships between the different components of this multi-layered notion: *conceptual/logical links*, and *process/causal links*. A mere list of relevant features is not enough, not even when complemented with frequency patterns.

More specifically, a satisfactory definition should be able to answer the following questions:

1. What are the relevant connections between the overall phenomenon of trust and its specific ingredients? Why are the latter within the former, and how does the former emerge from the latter?
2. What are the pair-wise relations between different features of trust? For instance, how do *belief* and *expectation*, or *outcome* and *reliance*, interact with each other?
3. What is the conceptual link and the process relationship between trust as attitude (*belief*, *evaluation*, *expectation*) and trust as decision and action (relying on, accepting, making oneself vulnerable, depending, etc.)?

1.2 Missed Components and Obscure Links

The content analysis of 72 definitions presented in the previous section reveals some relevant gaps in such definitions, as well as several notions that remain largely or completely implicit.

An aspect absolutely necessary but frequently ignored (or at least left unstated) is *the goal, the need,* relative to which and for the achievement of which the trustor counts upon the trustee.

[3] As we will discuss later on (Chapter 2, Section 2.4), we distinguish between *active trust* and *passive trust*. The former is related to the delegation of a positive action to Y, and to the expectation of obtaining the desired outcome from this action. The latter, instead, is just reduced to the expectation of receiving no harm from Y, no aggression: it is the belief that Y will not do anything dangerous for me, hence I do not need to be alerted, to monitor Y's behavior, to avoid something, to protect myself. This passive trust has a third, more primitive component: the idea or feeling that "there is nothing to worry about", "I am/feel safe with Y".