

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/Wooldridge: *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



A John Wiley and Sons, Ltd., Publication

This edition first published 2010
© 2010 John Wiley & Sons Ltd.,

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.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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 Yuri, Vania and Giulio, our realized dreams;
to the many colleagues who consciously or unconsciously 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

Foreword	xv
Introduction	1
1 Definitions of Trust: From Conceptual Components to the General Core	7
1.1 A Content Analysis	8
1.2 Missed Components and Obscure Links	12
1.3 Intentional Action and Lack of Controllability: Relying on What is Beyond Our Power	15
1.4 Two Intertwined Notions of Trust: Trust as Attitude vs. Trust as Act	17
1.5 A Critique of Some Significant Definitions of Trust	19
1.5.1 Gambetta: <i>Is Trust Only About Predictability?</i>	19
1.5.2 Mayer, Davis, & Schoorman: <i>Is Trust Only Willingness, for Any Kind of Vulnerability?</i>	19
1.5.3 McKnight: <i>The Black Boxes of Trust</i>	21
1.5.4 Marsh: <i>Is a Mere Expectation Enough for Modeling Trust?</i>	21
1.5.5 Yamagishi: <i>Mixing up the Act of Trusting and the Act of Cooperating</i>	22
1.5.6 Trust as Based on Reciprocity	26
1.5.7 Hardin: <i>Trust as Encapsulated Interest</i>	26
1.5.8 Rousseau: <i>What Kind of Intention is 'Trust'?</i>	30
References	31
2 Socio-Cognitive Model of Trust: Basic Ingredients	35
2.1 A Five-Part Relation and a Layered Model	36
2.1.1 <i>A Layered Notion</i>	36
2.1.2 <i>Goal State and Side Effects</i>	38
2.2 Trust as Mental Attitude: a Belief-Based and Goal-Based Model	38
2.2.1 <i>Trust as Positive Evaluation</i>	39
2.2.2 <i>The 'Motivational' Side of Trust</i>	44
2.2.3 <i>The Crucial Notion of 'Goal'</i>	45
2.2.4 <i>Trust Versus Trustworthiness</i>	47
2.2.5 <i>Two Main Components: Competence Versus Predictability</i>	47
2.2.6 <i>Trustworthiness (and trust) as Multidimensional Evaluative Profiles</i>	49

2.2.7	<i>The Inherently Attributional Nature of Trust</i>	50
2.2.8	<i>Trust, Positive Evaluation and Positive Expectation</i>	52
2.3	Expectations: Their Nature and Cognitive Anatomy	54
2.3.1	<i>Epistemic Goals and Activity</i>	54
2.3.2	<i>Content Goals</i>	55
2.3.3	<i>The Quantitative Aspects of Mental Attitudes</i>	56
2.3.4	<i>The Implicit Counterpart of Expectations</i>	58
2.3.5	<i>Emotional Response to Expectation is Specific: the Strength of Disappointment</i>	58
2.3.6	<i>Trust is not Reducible to a Positive Expectation</i>	60
2.4	'No Danger': Negative or Passive or Defensive Trust	60
2.5	Weakening the Belief-Base: Implicit Beliefs, Acceptances, and Trust by-Default	62
2.6	From Disposition to Action	64
2.6.1	<i>Trust That and Trust in</i>	66
2.6.2	<i>Trust Pre-disposition and Disposition: From Potential to Actual Trust</i>	67
2.6.3	<i>The Decision and Act of Trust Implies the Decision to Rely on</i>	69
2.7	Can we Decide to Trust?	72
2.8	Risk, Investment and Bet	73
2.8.1	<i>'Risk' Definition and Ontology</i>	74
2.8.2	<i>What Kinds of Taken Risks Characterize Trust Decisions?</i>	76
2.9	Trust and Delegation	77
2.9.1	<i>Trust in Different Forms of Delegation</i>	79
2.9.2	<i>Trust in Open Delegation Versus Trust in Closed Delegation</i>	80
2.10	The Other Parts of the Relation: the Delegated Task and the Context	82
2.10.1	<i>Why Does X Trust Y?</i>	82
2.10.2	<i>The Role of the Context/Environment in Trust</i>	83
2.11	Genuine Social Trust: Trust and Adoption	84
2.11.1	<i>Concern</i>	88
2.11.2	<i>How Expectations Generate (Entitled) Prescriptions: Towards 'Betrayal'</i>	88
2.11.3	<i>Super-Trust or Tutorial Trust</i>	89
2.12	Resuming the Model	91
	References	92
3	Socio-Cognitive Model of Trust: Quantitative Aspects	95
3.1	Degrees of Trust: a Principled Quantification of Trust	95
3.2	Relationships between Trust in Beliefs and Trust in Action and Delegation	97
3.3	A Belief-Based Degree of Trust	98
3.4	To Trust or Not to Trust: Degrees of Trust and Decision to Trust	101
3.5	Positive Trust is not Enough: a Variable Threshold for Risk Acceptance/Avoidance	107
3.6	Generalizing the Trust Decision to a Set of Agents	111

3.7	When Trust is Too Few or Too Much	112
3.7.1	<i>Rational Trust</i>	112
3.7.2	<i>Over-Confidence and Over-Diffidence</i>	112
3.8	Conclusions	114
	References	115
4	The Negative Side: Lack of Trust, Implicit Trust, Mistrust, Doubts and Diffidence	117
4.1	From Lack of Trust to Diffidence: Not Simply a Matter of Degree	117
4.1.1	<i>Mistrust as a Negative Evaluation</i>	118
4.2	Lack of Trust	119
4.3	The Complete Picture	120
4.4	In Sum	121
4.5	Trust and Fear	122
4.6	Implicit and by Default Forms of Trust	122
4.6.1	<i>Social by-Default Trust</i>	124
4.7	Insufficient Trust	125
4.8	Trust on Credit: The Game of Ignorance	126
4.8.1	<i>Control and Uncertainty</i>	126
4.8.2	<i>Conditional Trust</i>	127
4.8.3	<i>To Give or Not to Give Credit</i>	127
4.8.4	<i>Distrust as Not Giving Credit</i>	129
	References	131
5	The Affective and Intuitive Forms of Trust: The Confidence We Inspire	133
5.1	Two Forms of 'Evaluation'	134
5.2	The Dual Nature of Valence: Cognitive Evaluations Versus Intuitive Appraisal	134
5.3	Evaluations	135
5.3.1	<i>Evaluations and Emotions</i>	136
5.4	Appraisal	137
5.5	Relationships Between Appraisal and Evaluation	138
5.6	Trust as Feeling	140
5.7	Trust Disposition as an Emotion and Trust Action as an Impulse	141
5.8	Basing Trust on the Emotions of the Other	142
5.9	The Possible Affective Base of 'Generalized Trust' and 'Trust Atmosphere'	143
5.10	Layers and Paths	143
5.11	Conclusions About Trust and Emotions	144
	References	145
6	Dynamics of Trust	147
6.1	Mental Ingredients in Trust Dynamics	148
6.2	Experience as an Interpretation Process: Causal Attribution for Trust	150

6.3	Changing the Trustee's Trustworthiness	154
6.3.1	<i>The Case of Weak Delegation</i>	154
6.3.2	<i>The Case of Strong Delegation</i>	158
6.3.3	<i>Anticipated Effects: A Planned Dynamics</i>	161
6.4	The Dynamics of Reciprocal Trust and Distrust	164
6.5	The Diffusion of Trust: Authority, Example, Contagion, Web of Trust	168
6.5.1	<i>Since Z Trusts Y, Also X Trusts Y</i>	168
6.5.2	<i>Since X Trusts Y, (by Analogy) Z Trusts W</i>	173
6.5.3	<i>Calculated Influence</i>	173
6.6	Trust Through Transfer and Generalization	174
6.6.1	<i>Classes of Tasks and Classes of Agents</i>	175
6.6.2	<i>Matching Agents' Features and Tasks' Properties</i>	175
6.6.3	<i>Formal Analysis</i>	177
6.6.4	<i>Generalizing to Different Tasks and Agents</i>	178
6.6.5	<i>Classes of Agents and Tasks</i>	182
6.7	The Relativity of Trust: Reasons for Trust Crisis	184
6.8	Concluding Remarks	188
	References	189
7	Trust, Control and Autonomy: A Dialectic Relationship	191
7.1	Trust and Control: A Complex Relationship	191
7.1.1	<i>To Trust or to Control? Two Opposite Notions</i>	192
7.1.2	<i>What Control is</i>	192
7.1.3	<i>Control Replaces Trust and Trust Makes Control Superfluous?</i>	195
7.1.4	<i>Trust Notions: Strict (Antagonist of Control) and Broad (Including Control)</i>	196
7.1.5	<i>Relying on Control and Bonds Requires Additional Trust: Three Party Trust</i>	198
7.1.6	<i>How Control Increases and Complements Trust</i>	200
7.1.7	<i>Two Kinds of Control</i>	201
7.1.8	<i>Filling the Gap between Doing/Action and Achieving/Results</i>	203
7.1.9	<i>The Dynamics</i>	204
7.1.10	<i>Control Kills Trust</i>	205
7.1.11	<i>Resuming the Relationships between Trust and Control</i>	206
7.2	Adjusting Autonomy and Delegation on the Basis of Trust in Y	206
7.2.1	<i>The Notion of Autonomy in Collaboration</i>	209
7.2.2	<i>Delegation/Adoption Theory</i>	209
7.2.3	<i>The Adjustment of Delegation/Adoption</i>	213
7.2.4	<i>Channels for the Bilateral Adjustments</i>	222
7.2.5	<i>Protocols for Control Adjustments</i>	223
7.2.6	<i>From Delegation Adjustment to Autonomy Adjustment</i>	225
7.2.7	<i>Adjusting Meta-Autonomy and Realization-Autonomy of the Trustee</i>	225
7.2.8	<i>Adjusting Autonomy by Modifying Control</i>	226
7.2.9	<i>When to Adjust the Autonomy of the Agents</i>	227
7.3	Conclusions	230
	References	232

8	The Economic Reductionism and Trust (Ir)rationality	235
8.1	Irrational Basis for Trust?	236
8.1.1	<i>Is Trust a Belief in the Other's Irrationality?</i>	236
8.2	Is Trust an 'Optimistic' and Irrational Attitude and Decision?	239
8.2.1	<i>The Rose-Tinted Glasses of Trust</i>	239
8.2.2	<i>Risk Perception</i>	246
8.3	Is Trust Just the Subjective Probability of the Favorable Event?	247
8.3.1	<i>Is Trust Only about Predictability? A Very Bad Service but a Sure One</i>	247
8.3.2	<i>Probability Collapses Trust 'that' and 'in'</i>	248
8.3.3	<i>Probability Collapses Internal and External (Attributions of) Trust</i>	248
8.3.4	<i>Probability Misses the Active View of Trust</i>	250
8.3.5	<i>Probability or Plausibility?</i>	250
8.3.6	<i>Probability Reduction Exposes to Eliminative Behavior: Against Williamson</i>	250
8.3.7	<i>Probability Mixes up Various Kinds of Beliefs, Evaluations, Expectations about the Trustee and Their Mind</i>	252
8.4	Trust in Game Theory: from Opportunism to Reciprocity	254
8.4.1	<i>Limiting Trust to the Danger of Opportunistic Behavior</i>	255
8.4.2	<i>'To Trust' is not 'to Cooperate'</i>	255
8.5	Trust Game: A Procuste's Bed for Trust Theory	256
8.6	Does Trust Presuppose Reciprocity?	258
8.7	The Varieties of Trust Responsiveness	260
8.8	Trusting as Signaling	260
8.9	Concluding Remarks	261
	References	261
9	The Glue of Society	265
9.1	Why Trust is the 'Glue of Society'	265
9.2	Trust and Social Order	266
9.2.1	<i>Trust Routinization</i>	268
9.3	How the Action of Trust Acquires the Social Function of Creating Trust	268
9.4	From Micro to Macro: a Web of Trust	270
9.4.1	<i>Local Repercussions</i>	270
9.4.2	<i>Trans-Local Repercussions</i>	271
9.5	Trust and Contracts	272
9.5.1	<i>Do Contracts Replace Trust?</i>	272
9.5.2	<i>Increasing Trust: from Intentions to Contracts</i>	272
9.5.3	<i>Negotiation and Pacts: Trust as Premise and Consequence</i>	275
9.6	Is Trust Based on Norms?	275
9.6.1	<i>Does Trust Create Trust and does There Exist a Norm of Reciprocating Trust?</i>	277
9.7	Trust: The Catalyst of Institutions	278
9.7.1	<i>The Radical Trust Crisis: Institutional Deconstruction</i>	279
	References	279

10	On the Trustee's Side: Trust As Relational Capital	281
10.1	Trust and Relational Capital	282
10.2	Cognitive Model of Being Trusted	284
10.2.1	<i>Objective and Subjective Dependence</i>	285
10.2.2	<i>Dependence and Negotiation Power</i>	289
10.2.3	<i>Trust Role in Dependence Networks</i>	292
10.3	Dynamics of Relational Capital	297
10.3.1	<i>Increasing, Decreasing and Transferring</i>	297
10.3.2	<i>Strategic Behavior of the Trustee</i>	300
10.4	From Trust Relational Capital to Reputational Capital	301
10.5	Conclusions	302
	References	302
11	A Fuzzy Implementation for the Socio-Cognitive Approach to Trust	305
11.1	Using a Fuzzy Approach	306
11.2	Scenarios	306
11.3	Belief Sources	307
11.4	Building Belief Sources	307
11.4.1	<i>A Note on Self-Trust</i>	309
11.5	Implementation with Nested FCMs	310
11.6	Converging and Diverging Belief Sources	311
11.7	Homogeneous and Heterogeneous Sources	312
11.8	Modeling Beliefs and Sources	312
11.9	Overview of the Implementation	313
11.9.1	<i>A Note on Fuzzy Values</i>	315
11.10	Description of the Model	316
11.11	Running the Model	316
11.12	Experimental Setting	317
11.12.1	<i>Routine Visit Scenario</i>	317
11.12.2	<i>Emergency Visit Scenario</i>	319
11.12.3	<i>Trustfulness and Decision</i>	320
11.12.4	<i>Experimental Discussion</i>	321
11.12.5	<i>Evaluating the Behavior of the FCMs</i>	322
11.12.6	<i>Personality Factors</i>	322
11.13	Learning Mechanisms	323
11.13.1	<i>Implicit Revision</i>	324
11.13.2	<i>Explicit Revision</i>	324
11.13.3	<i>A Taxonomy of Possible Revisions</i>	325
11.14	Contract Nets for Evaluating Agent Trustworthiness	326
11.14.1	<i>Experimental Setting</i>	326
11.14.2	<i>Delegation Strategies</i>	327
11.14.3	<i>The Contract Net Structure</i>	328
11.14.4	<i>Performing a Task</i>	329
11.14.5	<i>FCMs for Trust</i>	329
11.14.6	<i>Experiments Description</i>	330
11.14.7	<i>Using Partial Knowledge: the Strength of a Cognitive Analysis</i>	333

11.14.8	<i>Results Discussion</i>	339
11.14.9	<i>Comparison with Other Existing Models and Conclusions</i>	341
	References	342
12	Trust and Technology	343
12.1	Main Difference Between Security and Trust	344
12.2	Trust Models and Technology	345
12.2.1	<i>Logical Approaches</i>	346
12.2.2	<i>Computational Approach</i>	347
12.2.3	<i>Different Kinds of Sources</i>	347
12.2.4	<i>Centralized Reputation Mechanisms</i>	348
12.2.5	<i>Decentralized Reputation Mechanisms</i>	349
12.2.6	<i>Different Kinds of Metrics</i>	350
12.2.7	<i>Other Models and Approaches to Trust in the Computational Framework</i>	351
12.3	Concluding Remarks	354
	References	354
13	Concluding Remarks and Pointers	359
13.1	Against Reductionism	359
13.2	Neuro-Trust and the Need for a Theoretical Model	360
13.3	Trust, Institutions, Politics (Some Pills of Reflection)	361
13.3.1	<i>For Italy (All'Italia)</i>	362
	References	363
Index		365

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