TRUST THEORY

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

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

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