Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Research Report
RR-00-02

Vertrauen und Betrug in Multi-Agenten Systemen

Erweiterung des Vertrauensmodells von Castelfranchi und Falcone um eine Kommunikationskomponente

Michael Schillo

Januar 2000

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Postfach 2080
67608 Kaiserslautern, FRG
Tel: +49 (631) 205-3211
Fax: +49 (631) 205-3210
E-Mail: info@dfki.uni-kl.de

Stuhlsatzenhausweg 3
66123 Saarbrücken, FRG
Tel: +49 (631) 302-5252
Fax: +49 (631) 302-5341
E-Mail: info@dfki.de

WWW: http://www.dfki.de

Deutsches Forschungszentrum für Künstliche Intelligenz

DFKI GmbH

German Research Center for Artificial Intelligence

Founded in 1988, DFKI today is one of the largest non-profit contract research institutes in the field of innovative software technology based on Artificial Intelligence (AI) methods. DFKI is focusing on the complete cycle of innovation — from world-class basic research and technology development through leading-edge demonstrators and prototypes to product functions and commercialisation.

Based in Kaiserslautern and Saarbrücken, the German Research Center for Artificial Intelligence ranks among the important "Centers of Excellence" world-wide.

An important element of DFKI's mission is to move innovations as quickly as possible from the lab into the marketplace. Only by maintaining research projects at the forefront of science can DFKI have the strength to meet its technology transfer goals.

DFKI has about 115 full-time employees, including 95 research scientists with advanced degrees. There are also around 120 part-time research assistants.

Revenues for DFKI were about 24 million DM in 1997, half from government contract work and half from commercial clients. The annual increase in contracts from commercial clients was greater than 37% during the last three years.

At DFKI, all work is organised in the form of clearly focused research or development projects with planned deliverables, various milestones, and a duration from several months up to three years.

DFKI benefits from interaction with the faculty of the Universities of Saarbrücken and Kaiserslautern and in turn provides opportunities for research and Ph.D. thesis supervision to students from these universities, which have an outstanding reputation in Computer Science.

The key directors of DFKI are Prof. Wolfgang Wahlster (CEO) and Dr. Walter Olthoff (CFO).

DFKI's six research departments are directed by internationally recognised research scientists:

- Information Management and Document Analysis (Director: Prof. A. Dengel)
- Intelligent Visualisation and Simulation Systems (Director: Prof. H. Hagen)
- Deduction and Multiagent Systems (Director: Prof. J. Siekmann)
- Programming Systems (Director: Prof. G. Smolka)
- Language Technology (Director: Prof. H. Uszkoreit)
- Intelligent User Interfaces (Director: Prof. W. Wahlster)

In this series, DFKI publishes research reports, technical memos, documents (e.g. workshop proceedings), and final project reports. The aim is to make new results, ideas, and software available as quickly as possible.

Prof. Wolfgang Wahlster

Director

Vertrauen und Betrug in Multi-Agenten Systemen

Erweiterung des Vertrauensmodells von Castelfranchi und Falcone um eine Kommunikationskomponente

Michael Schillo

DFKI-RR-00-02

This work has been supported by the German National Scholarship Foundation (Studienstiftung des deutschen Volkes).

ISSN 0946-008X

Zusammenfassung

Diese Arbeit beschäftigt sich mit betrügerischen Agenten in Künstlichen Gesellschaften und damit, wie andere Agenten sich vor ihnen schützen können. Zu diesem Zweck werden Agenten mit Berechnungsmodellen für zwei Konzepte von „Vertrauen“ ausgestattet. Zum einen berechnen sie Vertrauen in Interaktionspartner mit einer präzisierten Variante des Modells von Castelfranchi und Falcone. Zum anderen benutzen sie eine hier vorgestellte Form von Vertrauen, um mit anderen über das Verhalten von unbekannten Agenten zu kommunizieren. Durch diesen Datenaustausch sind sie in der Lage, fremde Agenten wesentlich schneller und besser einzuschätzen. Mit diesem Wissen können sich Agenten effektiver vor betrügerischen und nicht-benevolenten Agenten schützen. Das Vertrauen in Kommunikationspartner schafft einen „sozialen Kitt“, über den innerhalb einer Gruppe Informationen zuverlässig ausgetauscht werden können.

Desweiteren wird hier das *Offen Gespielte Gefangenendilemma mit Partnerauswahl* vorgestellt. Dabei handelt es sich um ein spieltheoretisches Modell, in dem Agenten andere betrügen können. Diese Variation des Gefangenendilemmas dient als Experimentalumgebung für heterogene Agentengesellschaften. Diese Experimentalumgebung besitzt wichtige Eigenschaften von Anwendungsszenarien wie z.B. die Kooperation in Virtuellen Märkten. Sie ist so gestaltet, daß die Effektivität von Strategien im Umgang mit betrügerischen Agenten untersucht werden kann. Dies bedeutet, daß mit ihrer Hilfe Turniere, ähnlich dem in der Literatur viel beachteten Turnier von Axelrod, durchgeführt werden können. Schließlich wird diese Experimentalumgebung genutzt, um das hier vorgestellte Modell des Vertrauens in Kommunikationspartner in einer Reihe von Experimenten, in denen die Agenten kein *a priori* Wissen über das Verhalten anderer haben, zu analysieren. Bei dieser Analyse werden Konfigurationen von verschieden ehrlichen und kooperationswilligen Agenten untersucht.

In der Evaluation des Ansatzes zeigt sich, daß Agenten durch den Austausch von Wissen mit anderen vertrauenswürdigen Agenten ihre Interaktionspartner besser einschätzen können. Insbesondere sind sie in der Lage, Interaktionspartner einzuschätzen, die sie selbst noch nie beobachten konnten. In den untersuchten Agentengesellschaften bedeutet dies einen Performanzgewinn von mehr als fünfzehn Prozent, ohne daß die Agenten ein *a priori* Wissen über das Verhalten ihrer Interaktionspartner haben. Die Benutzung von Vertrauen und Kommunikation zahlt sich insbesondere dann aus, wenn nur wenige Beobachtungen über das Verhalten anderer zur Verfügung stehen.

Danksagung

Mein Dank gilt

- Herrn Prof. Siekmann für die Möglichkeit diese Arbeit an seinem Lehrstuhl durchführen zu können. Ihm und insbesondere den Lehrstuhlbewohnern Walter Bieniossek, Lassaad Cheikhrouhou, Christian Gerber, Christoph Jung und Martin Pollet danke ich für die angenehme Arbeitsatmosphäre und ihre Hilfsbereitschaft. Allen am Lehrstuhl gilt auch mein Dank für die Geduld während der Durchführung der rechenzeitintensiven Simulationen.
- in ganz besonderem Maße Petra Funk für ihre engagierte Betreuung und ihre Unterstützung (und natürlich dafür, daß sie sich sogar gegen besseres Wissen darauf eingelassen hat, daß diese Arbeit nicht mit TEX verfaßt wurde).
- Jürgen Lind für den Hinweis auf die Arbeiten von Stephen Marsh und natürlich für die „Lind-Tools".
- Michael Rovatsos für die vielen Diskussionen, die diese Arbeit wesentlich beeinflußt haben.
- Ralf Schäfer und Emil Weydert für die wertvollen Diskussionen zu Bayes'schen Netzen, Matthias Schunter für die Literaturhinweise aus dem Bereich der Kryptographie.
- Der Studienstiftung des deutschen Volkes für die großzügige Unterstützung während meines Studiums.
- Sandra Betrand, Nina Koch, Andreas Meier und Markus Utesch für ihr Interesse und die Geduld bei der Bearbeitung des Manuskripts.
- Nicht zuletzt Jessica Seibert, der ich dafür dankbar bin, daß sie mir mathematisch mit Rat und Tat geholfen hat, sich des Manuskripts angenommen hat und ganz besonders danke ich für ihr - Vertrauen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Definitionsverzeichnis

Kapitel 1
Einleitung

„I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Alan Turing, 1950

„Heute wissen wir, daß die Geistmetapher in der Informatik eine Paradigmenrevolution auslöste, die zu weitreichenden technischen Innovationen geführt hat. Der soziologisch springende Punkt aber liegt in der Frage: Kann die sozionische Forschung irgendwann einen ähnlich bahnbrechenden computerreferenziellen Innovationsbeitrag leisten...?"

Thomas Malsch, 1996

Nachdem die Erforschung der Künstlichen Intelligenz in den letzten Jahrzehnten eine enge Kooperation mit der Psychologie eingegangen ist, steht eine ähnlich intensive Kontaktaufnahme mit der Soziologie bevor. In diesem Kontext steht die vorliegende Diplomarbeit, die soziologische und psychologische Erkenntnisse für ein Problem der Verteilten Künstlichen Intelligenz nutzt. Dieses Kapitel gibt eine kurze Einführung in die Ziele und Anwendungen dieser Arbeit und erläutert die Gliederung der folgenden Kapitel.

1.1. Problemstellung

Wie findet ein Agent heraus, ob ein anderer Agent ihn betrügen will? Rosenschein und Genesereth haben festgestellt, daß die Annahme, in einem *Offenen System* (wie beispielsweise dem Internet) seien alle Akteure benevolent, ebenso weit verbreitet wie unrealistisch ist (Rosenschein und Genesereth, 1985). Selbst bis heute gibt es wenige Konzepte, wie autonome Agenten sich ohne diese Annahme verhalten sollen (Marsh, 1994). Armstrong und Durfee schlagen zwei mögliche Wege vor. Zum einen könnte eine Art „Polizei" eingeführt werden, deren Aufgabe es ist, betrügerische

Agenten zu entfernen. Damit würden die Gefahren durch dauerhaft betrügerische Agenten zwar verringert, aber nicht ausgeschlossen. Für Agenten, die sich nicht selbst schützen können, entstehen die selben Probleme wie in der menschlichen Gesellschaft: Ein „Polizeiagent" kann oft erst eingreifen, wenn schon ein Schaden entstanden ist. Der Mechanismus greift also für den Geschädigten möglicherweise zu spät. Deshalb ziehen Armstrong und Durfee eine weitere Möglichkeit vor. Diese besteht darin, Agenten mit der Fähigkeit auszustatten, das Verhalten anderer auf seine Gefährlichkeit hin einzuschätzen und gegebenenfalls die Interaktion zu verweigern (Armstrong und Durfee, 1998). Insbesondere wenn situierte Agenten für ihren Benutzer eine Aufgabe unbeaufsichtigt erledigen, sollten sie in der Lage sein, gefährliche Agenten zu erkennen.

Dieses Erkennen beruht notwendigerweise auf der Auswertung von Verhalten aus der Vergangenheit. Was aber ist, wenn einem Agenten nur wenig oder gar keine Information zur Verfügung steht? Künstliche Gesellschaften, wie z.B. Multi-Agenten Systeme, sind nach Les Gasser per Definition *sozial* und es liegt nahe, Gesellschaftswissenschaften wie die Soziologie und die Sozialpsychologie zu Rate zu ziehen (Gasser, 1991). In diesen Wissenschaften ist es ein beobachtetes Phänomen, daß in Situationen, in denen Informationen über das Verhalten andere nicht zugänglich sind, die Aussagen und Einschätzungen Dritter benutzt werden. Dabei spielt das *Vertrauen*, das in diese „Informanten" gesetzt wird, eine zentrale Rolle. An dieser Stelle betritt die Verteilte Künstliche Intelligenz Neuland. Sogar in so ausgeklügelten Systemen wie COMRIS (CO-habited Mixed-Reality Information Spaces), die die soziale Interaktion ihrer Agenten auf Emotionen aufbauen, findet Vertrauen (noch) keine Beachtung (Conamero und Van de Velde, 1997). Stehen Beobachtungen und Aussagen über andere Agenten zur Verfügung, so folgt aus Arbeiten der Soziologie und Sozialpsychologie, daß ein wesentliches Kriterium zur Bewertung des Verhaltens ist, ob Werte und Normen eingehalten wurden.

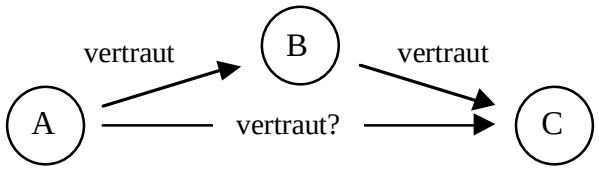

Abbildung 1: Ziel der Modellierung: Vertrauen in unbekannte Dritte

Derartige Erkenntnisse werden in dieser Arbeit genutzt, um autonomen Agenten in einem Offenen System einen Abwehrmechanismus im Sinne von Armstrong und Durfee zu geben. Wir zeigen, wie Agenten die Kommunikation mit anderen benutzen können, um ihre Einschätzungen von Interaktionspartnern zu optimieren und schneller zu erhalten. Exemplarisch ist dies in Abbildung 1 gezeigt: hat Agent *A* mit *B* über *C* kommuniziert, kann er aus seiner Einschätzung von *B* eine Einschätzung von *C* ableiten. Unser Vorgehen berücksichtigt, daß Kommunikation mit Agenten stattfindet, die möglicherweise selbst nicht vertrauenswürdig sind und benutzt auch deren Informationen zur Evaluation von Agenten. Um dies zu erreichen, beschreiben wir Vertrauen nicht in zweiwertiger Logik, sondern definieren es als eine Wahrscheinlichkeit. Mit dieser Vorgehensweise können verschieden Grade von Vertrauen betrachtet werden, ohne daß Informationen unter einem Schwellenwert nutzlos werden. Um die Entscheidungen für oder gegen einen Interaktionspartner zu

treffen, wird das Vertrauensmodell von Castelfranchi und Falcone benutzt (1998; 1999). Zentral für dieses Modell ist eine Approximation der Intentionen des möglichen Interaktionspartners. Diese Approximation wird aber bei Castelfranchi und Falcone als bereits berechnet vorausgesetzt. Die Autoren geben nicht an, wie dieser Wert zustande kommt. Im folgenden wird ein Algorithmus für seine Berechnung angegeben. Der Algorithmus verwendet zwei verschieden Ressourcen: Zum einen werden die selbst gemachten Beobachtungen als zuverlässige Datenquelle genutzt. Zum anderen werden, falls die eigenen Beobachtungen veraltet oder nicht ausreichend sind, die Aussagen von anderen Agenten benutzt. Insbesondere die Verwendung solcher Aussagen ist in *Offenen Systemen* mit nicht-benevolenten Agenten eine Herausforderung.

1.2. Anwendungsgebiet

In vielen Domänen müssen egoistische Akteure (z.B. Firmen) kooperieren, um Aufgaben effizient, kostensparend und sicher zu lösen. Dabei ist es nicht immer möglich, alle Akteure selbst zu kennen, da z.B. die Anzahl der Akteure zu groß ist, man selbst erst neu im Metier ist, sich die Zusammensetzung der Gesellschaft schnell verändert etc. Gerade in einer solchen Situation reicht es nicht aus, daß sich Agenten auf das wohlwollende Verhalten anderer Agenten verlassen. Das rapide Wachstum des Internet und die zunehmende Computerisierung tragen ebenfalls dazu bei, daß dieses Problem enorme Beachtung findet (z.B. (Sandholm und Lesser, 1995), (Zeng und Sycara, 1996)). Um das Verhalten von Agenten einzuschätzen, ist Vertrauen ein Mechanismus, der in der Literatur immer mehr als Approximationsmethode für ein solches Problem behandelt wird ((Brainov und Sandholm 1999), (Tan und Thoen 1999)). Diese Approximation kann bei geeigneten Berechnungen beschleunigt und verbessert werden, wenn Erfahrungen von anderen Agenten zur Verfügung stehen. Bei genauer Betrachtung zeigt sich, daß Vertrauen fester Bestandteil jeder Art von Interaktion ist. Diese Tatsache wurde jedoch in bisherigen Implementierungen von Multi-Agenten Systemen ignoriert, indem Vertrauen implizit vorausgesetzt wurde (Marsh, 1994). In Offenen Systemen wie dem Internet kann diese Annahme nicht mehr gemacht werden.

Bei der Automatisierung der Koordination durch Verhandlungsagenten kommt Vertrauen in andere Agenten (da nicht von benevolenten Agenten ausgegangen werden kann) und der Einschätzung ihres Verhalten große Bedeutung zu. Das eigene Wohlergehen oder der eigene Profit hängen davon ab, wie gut diese Einschätzungen sind. Es gibt eine Reihe von Szenarien, in denen es wichtig ist, die Daten (auf denen Vertrauen dann aufbauen kann) möglichst schnell zu erhalten und das Vertrauensmodell so schnell wie möglich zu einem guten Wert konvergieren zu lassen. Solche Szenarien sind z.B.:

1. Der Schutz von Agenten, die von Rechner zu Rechner migrieren (mobile Agenten).
2. Firmenkooperation in virtuellen Märkten (*Electronic Commerce).*
3. Das Verwalten von öffentlichen Schlüsseln in kryptographischen Systemen.
4. Das Routen von Nachrichten über vertrauenswürdige Internetrechner.
5. Kooperation beim *Information Retrieval* im Internet.

Auf diese Szenarien und welchen Nutzen Kommunikation bezüglich der Vertrauenswürdigkeit von anderen hat, wird in Kapitel 3 genauer eingegangen.

Die Einschätzung des Verhaltens von Agenten aufgrund von Aussagen Dritter ist keine leichte Aufgabe. Es muß berücksichtigt werden, daß diese Agenten eigene Motive und Intentionen haben. Ein Agent, der sich auf andere verläßt, muß also zwei Arten von Unsicherheit bewältigen können: Zum einen die mögliche Unvollständigkeit und die Nichtrepräsentativität der Daten. Zum anderen (und dies hat weitreichendere Folgen) muß er mit der absichtlichen Manipulation der übermittelten Daten durch den Sender umgehen können. Das Problem kann also nicht allein durch einen probabilistischen Ansatz bezüglich einer Stichprobe aus einer gleichverteilten Datenmenge gelöst werden, sondern erfordert eine Analyse des Verhaltens und der Motive des Senders.

1.3. Ergebnisse

Der vorgeschlagene Ansatz behandelt dieses Problem erfolgreich. In einer Reihe von Simulationen zeigt sich, daß Agenten mit diesem Ansatz bessere Modelle des Verhaltens anderer erstellen und die Adaption ihrer Interaktionen für die Agenten eine Performanzsteigerung bewirkt. Diese Performanzsteigerung wird in verschiedener Hinsicht überprüft. Es wird untersucht, wie sich die Performanz verhält, wenn mehr oder weniger Kommunikation erlaubt ist und wenn mehr bzw. weniger Daten selbst beobachtet werden können. Es zeigt sich, daß der Ansatz deutliche Vorteile bringt, je weniger eigene Beobachtung möglich ist. Darüber hinaus ist die Performanzsteigerung umso größer, je mehr Datenaustausch (also Kommunikation) möglich ist. Es stellt sich außerdem heraus, daß schon sehr wenig Kommunikation deutliche Verbesserung der Performanz bewirkt. Das ist ein Hinweis darauf, daß selbst in Szenarien mit hohen Kommunikationskosten der vorgeschlagene Ansatz gute Resultate liefert. Auf die Analyse wird in Kapitel 6 detailliert eingegangen.

1.4. Aufbau der Arbeit

Im folgenden Kapitel 2 geben wir einen Überblick über die relevante Literatur und legen dabei das Fundament, auf dem die anderen Kapitel aufbauen. Zunächst behandeln wir den Bereich der Multi-Agenten Systeme, da hier der Schwerpunkt der Diplomarbeit liegt. Es folgt eine Übersicht über Vertrauen aus psychologischer und soziologischer Sicht. Das Modell von Castelfranchi und Falcone, welches im weiteren benutzt wird, stellen wir ebenfalls vor. Mit einer kurzen Zusammenfassung der später genutzten Begriffe aus der Entscheidungs- und Spieltheorie schließen dieses Kapitel ab.

Als Fortführung der Einleitung erläutern wir in Kapitel 3 die Problemstellung detailliert und zeigen den Umfang der Anwendungsgebiete auf. Die vorgeschlagene Lösung legen wir in den beiden darauf folgenden Kapiteln dar. In Kapitel 4 beschreiben wir die Lösung formal, entwickeln für die Berechnung notwendige Gleichungen und schlagen eine Experimentalumgebung vor, in der die Lösung getestet und gegebenenfalls im Popper'schen Sinne falsifiziert werden kann. In Kapitel 5 beschreiben wir den praktischen Teil der vorliegenden Arbeit. Dies umfaßt

sowohl die Implementierung des beschriebenen Formalismus, als auch die Realisierung der Experimentalumgebung. Nachdem die Methode und ihre Anforderungen vollständig präsentiert wurden, folgt eine Rechtfertigung der beschriebenen Vorgehensweise. Die Ergebnisse der Evaluation des Formalismus und seiner Implementierung stellen wir in Kapitel 6 vor. Diese Analyse zeigt, in welchem Maße Agenten von der Lösung profitieren und in welchen Situationen der Effekt besonders signifikant ist.

Schließlich fassen wir in Kapitel 7 die Arbeit und ihre Resultate noch einmal zusammen. Außerdem beschreiben wir Anknüpfungspunkte für die Fortsetzung der hier beschriebenen Studien. In den Anhängen befinden sich die der Analyse zugrundeliegenden Daten, eine Beschreibung des entwickelten Programms und die Konfigurationen, die für die Analyse benutzt wurden.

Kapitel 2
Vertrauen und Künstliche Gesellschaften in der Forschung: Ein Überblick

„DAI systems, as they involve multiple agents, are *social* in character."

Les Gasser, 1990.

In diesem Kapitel wollen wir die theoretische Grundlage für die vorliegende Arbeit erläutern. Wir werden zunächst einen Abriß über die wichtigsten Konzepte aus dem Bereich der Multi-Agenten Systeme (MAS) vorstellen, der sich im wesentlichen an Müller (1996) und Weiß (1996) anlehnt. Im zweiten Abschnitt diskutieren wir das Verhältnis der Verteilten Künstlichen Intelligenz Forschung zur Soziologie und gehen auf den Begriff „sozial" ein. Dieser Abschnitt basiert im wesentlichen auf Malsch et al. (1996). Im dritten Abschnitt betrachten wir eine Gegenüberstellung verschiedener Definitionsansätze des Konzeptes *Vertrauen* und stellen die Unterschiedlichkeit der möglichen Perspektiven heraus. Wichtigste Texte für den interdisziplinären Teil sind von Marsh (1994) und Deutsch (1973). In einem weiteren Abschnitt wird das Vertrauensmodell von Castelfranchi und Falcone (1998) vorgestellt. Mit einer kurzen Behandlung der benötigten Begriffe der Entscheidungs- und Spieltheorie runden wir den Überblick ab.

2.1. VKI und Multi-Agenten Systeme

Die Forschungsrichtung der Verteilten Künstlichen Intelligenz (VKI, Verteilte KI), oder *Distributed Artificial Intelligence* (DAI) wie sie im englischen heißt, ist ein Teilbereich der Forschungsrichtung *Künstliche Intelligenz* (KI). In diesem Systemparadigma wird davon ausgegangen, daß es komplexe Probleme gibt, die von einer Vielzahl kleinerer Einheiten besser gelöst werden können als von einem einzigen *Problemlöser*. Minsky beschreibt in seinem Buch „Society of Mind" (1986) eine interessante Verbindung zwischen Verteilter Künstlicher Intelligenz zur Kognitionswissenschaft:

Er legt dort seine Ansicht dar, daß das menschliche Denken aus dem Zusammenwirken einer „Gesellschaft“ von Problemlösern besteht, die bei der Realisierung ihrer Lösungsvorschläge in Konkurrenz stehen. Die Betrachtung eines Problems unter dem Aspekt der Verteilung kann in vielerlei Hinsicht von Vorteil sein.

Erstens kann es sein, daß das für eine Problemlösung benötigte Wissen von sich aus räumlich verteilt ist, wie dies z.B. bei der Flugleitkontrolle internationaler Flughäfen der Fall ist. Zweitens können für die Problemlösung Fähigkeiten notwendig sein, über die ein einzelner Akteur nicht verfügt. Drittens verspricht sich die Forschung von der Modellierung von Einheiten, die über spezielles Wissen und Fähigkeiten verfügen, eine bessere Modularisierung, Flexibilität und kürzere Reaktionszeiten (Chaib-Draa, 1994). Die VKI beschäftigt sich daher auch mit der Verteilung der Intelligenz einer Agentengesellschaft auf mehrere aktive Problemlöseeinheiten. Dabei berücksichtigt sie eventuell unvollständiges Wissen über die Welt und die dadurch auftretenden Probleme der Interaktion ((Müller, 1993), (Durfee, 1991)). Les Gasser beschreibt sechs Prinzipien für die VKI. Für ihn ist zunächst die Existenz und Interaktion mehrerer Agenten ein fundamentaler Aspekt der VKI. Diese Wissenschaft soll die Spannung zwischen dem situierten, dem lokalen und dem pragmatischen Charakter von Handeln und Wissen behandeln. Weiterhin müssen Wissensrepräsentation und gemeinsames Schließen von verteilter Repräsentation und verteiltem Schließen ausgehen. Alle Agenten sind dabei in ihren Ressourcen beschränkt. Außerdem ist für die VKI von zentralem Interesse, wie Gruppen von Agenten ihre Handlungen miteinander koordinieren können, so daß das gemeinsame Handeln trotz nicht-deterministisch auftretenden Fehlern und Inkonsistenzen robust und andauernd ist (Gasser, 1991). Carl Hewitt hat mit seinen *Open Information System Semantics* (*OISS*) versucht, ein gemeinsames Vokabular und eine eigene Methodologie für diese Forschungsrichtung zu entwerfen. Dabei steht der Begriff der *Offenen Systeme* für die Unberechenbarkeit der Menge aller möglichen Zustände und Ereignisse in diesen Systemen ((Hewitt und de Jong, 1984),(Hewitt, 1991), eine Kritik der *OISS* findet sich in (Gasser, 1991)).

Traditionell unterscheidet man in der Verteilten KI zwei Arten von Systemen: verteilte Problemlösesysteme und Multi-Agenten Systeme ((Bond und Gasser, 1988), (Durfee und Rosenschein, 1994), (von Martial, 1992)). Verteilte Problemlösesysteme sind jeweils für ein spezielles Problem konzipiert. Dazu wird ein Problem in Teilaufgaben zerlegt und für diese Teilaufgaben werden spezielle Problemlöser entworfen. Kooperation und Koordination sind in einem solchen System also explizit vorgegeben.

Die Teildidsziplin der Multi-Agenten Systeme (MAS) beschäftigt sich zum einen mit dezentralem Problemlösen durch Agenten, die generischer und daher auch autonomer als die verteilten Problemlöser sind. Eine zentrale Kontrolle der Berechnung gibt es hier nicht. Bond und Gasser schreiben hierzu:

> *Multi-Agenten Systeme beschäftigen sich mit der Koordination intelligenten Verhaltens einer Anzahl von autonomen, intelligenten Agenten: der Aufgabe, wie sie ihr Wissen, ihre Ziele, Fähigkeiten und Pläne vereint einsetzen können um Probleme zu lösen.“* *(Bond und Gasser, 1988)*

Dies ergänzt die erste der drei Fragestellungen von Müller, mit denen sich die Multi-Agenten System Forschung definiert. Seine erste Fragestellung lautet: „Was ist ein Agent?“ und geht darauf ein, welche Eigenschaften ein Agent besitzen kann und welche er besitzen muß. An dieser Stelle findet ein großer Transfer von Wissen und Methoden von der traditionellen KI in die VKI statt.

Zum anderen beschäftigt sich die VKI auch mit dem Studium quasi-soziologischer Prozesse. Nach Müller beschäftigt sich die Multi-Agenten System Forschung auch mit der Frage, wie sich eine Agentengesellschaft organisiert. Denn es entsteht eine Reihe von Problemen, wenn eine Agentengesellschaft entworfen wird. Es muß eine Möglichkeit zum Handeln (oder aber zumindest eine minimale Form der Kommunikation) vorgegeben werden und es muß erforscht werden, welche Interaktion dem Gesamtverhalten dienlich ist. Müller spricht vorsichtig vom Begriff der *Agentensoziologie*. Als dritte Fragestellung sieht Müller die Eingrenzung der Menge von Problemen, die eine Agentengesellschaft lösen kann. Die Forschungsrichtung soll klären, bei welchen Problemen welche Agentengesellschaft eingesetzt werden sollte. Außerdem sind Erkenntnisse darüber erwünscht, welche Zusammenhänge zwischen Gesellschaftsform und der Güte der Problemlösung bestehen.

2.1.1. Agenteneigenschaften und Agentenarchitekturen

Da die VKI noch eine sehr junge Disziplin ist, existieren zahlreiche Definitionen der Hauptbegriffe nebeneinander (Müller, 1993). Es wird sicherlich noch einige Zeit dauern, bis sich eine einheitliche Begriffsbildung über die definierenden Eigenschaften eines Agenten bzw. eines Multi-Agenten Systems entwickelt hat. Die grundlegendste und am weitesten akzeptierte Definition ist die von Russell und Norvig:

> *„Ein* Agent *ist eine Einheit, von der man sagen kann, daß sie ihre* Umwelt *über* Sensoren wahrnimmt *und sie mit Hilfe von* Effektoren beeinflußt.*“ (Russell und Norvig, 1996)*

Trotz der Meinungsverschiedenheiten über eine einheitliche Definition haben sich bestimmte zentrale Begriffe herausgebildet. Eine Reihe von Autoren ((Wooldridge und Jennings, 1995), (Green et al., 1997), (Weiß, 1996)) haben folgende Eigenschaften zur Charakterisierung von Agenten beschrieben:

- *Autonomie*: Agenten arbeiten ohne direkte Intervention durch Menschen und besitzen selbst die Kontrolle über ihr Handeln.
- *Soziale Fähigkeiten*: Agenten kommunizieren mit anderen Agenten mittels einer Agenten-Kommunikationssprache und versuchen Lösungen gemeinsamer Probleme kooperativ zu planen.
- *Reaktivität*: Agenten nehmen ihre Umwelt wahr und reagieren rechtzeitig und angemessen auf deren Veränderungen.
- *Pro-Aktivität:* Agenten agieren nicht nur als Antwort auf Veränderungen in ihrer Umwelt, sondern sind in der Lage auf zielgerichtete Art und Weise selbst die Initiative zu ergreifen.

Dies bezeichnen Wooldridge und Jennings als *schwache Agenteneigenschaft* (*weak notion of agency*). Die *starke Agenteneigenschaft* (*strong notion of agency*) beinhaltet für sie mentalistische Begriffe wie Wissen, Glaube, Intention und Verpflichtung (Wooldridge und Jennings, 1995). Weiß fügt den notwendigen Eigenschaften für einen Agenten noch Rationalität, Mobilität, Introspektion, Wahrhaftigkeit und Benevolenz hinzu. Funk und Lind betonen die Wichtigkeit des Begriffs des Wissens über sich selbst und andere (Funk und Lind, 1997). Sie argumentieren, daß ohne das Wissen über sich selbst die übrigen Punkte nur schwer zu realisieren sind und dieses damit in das Repertoire eines Agenten aufgenommen werden müßte.

Shoham beschrieb 1993 wie das Verhalten eines Agenten für verschiedene Situationen modelliert werden kann. Er führte das Paradigma des *Agenten-orientierten Programmierens* (*AOP*) ein (Shoham, 1993). Er erreicht dies mit Hilfe der Programmiersprache *Agent-0*. Mit ihr können Regeln aufgestellt werden, die Zustände und Zustandsänderungen für das Verhalten des Agenten bestimmen. Für Shoham waren die Begriffe Autonomie, Kommunikation, Wissen, Fähigkeiten und Verpflichtungen (*commitments, s. u.*) die Konzepte, die den Begriff *Agent* definieren. In seiner Syntax und Semantik interner Zustände beschreibt er (an die Prädikaten- und Temporallogik angelehnt) den Zusammenhang zwischen Glauben, Verpflichtung und Fähigkeit und wie eine Handlungsplanung erfolgen kann. *AOP* ist eine Spezialisierung der objektorientierten Programmierung, da die Objekte zusätzlich über interne Zustände verfügen, mit deren Hilfe die oben angegebenen Eigenschaften wie Autonomie etc. modelliert werden können. Agenten sind also eine Teilmenge der Objekte und somit sollte der Agentenbegriff deutlich von dem Modul- und dem Objektbegriff aus dem Software Engineering unterschieden werden (Müller, 1993).

Beim Realisieren von Agenten gibt es zwei prinzipiell unterschiedliche Konzepte: Agenten können *reaktiv* oder *deliberativ* sein. *Reaktiv* (auch *non-deliberativ* oder *verhaltensbasiert*) bedeutet für einen Agenten, daß er seine Entscheidungen aufgrund von Stimulus-Response-Paaren trifft (Brooks, 1991). Vorbilder für reaktive Systeme sind nach Sundermeyer z.B. Insekten (Sundermeyer, 1993). Deliberativ (*reflektiv*) bedeutet, daß der Agent über eine eigene Wissensrepräsentation verfügt, die explizit Wissen über das Umfeld speichert. Weiterhin besitzt der Agent ein Planungssystem, das seine weiteren Aktionen auswählt ((Shoham, 1993), (Cohen und Levesque, 1985)). Die deliberative Auffassung von einem Agenten setzt die Tradition der klassischen KI und deren symbolverarbeitenden Ansätze fort. Vorbilder für solche Systeme sind menschliche Akteure (Müller, 1993). Eine der prominentesten deliberativen Agentenarchitekturen ist die sogenannte *BDI*-Architektur, deren Konzeption von Bratman, Israel und Pollack stammt und aus einem Projekt mit Georgeff, Konolige, Cohen, Hayes und Lansky hervorging. *BDI* steht für die drei Begriffe *belief*, *desire* und *intention*.

- **Beliefs** (Fakten, die geglaubt werden) sind dabei Annahmen eines Agenten über den momentanen Zustand der Welt und ihre voraussichtliche weitere Veränderung. Meist werden sie mit Hilfe von „mögliche Welten" Semantiken beschrieben, die für jede Situation definieren, welche Weltzustände nach Ansicht des Agenten in der Zukunft daraus hervorgehen können.

- **Desires** (Wünsche) sind, abstrakt gesehen, die Zustände, die ein Agent für wünschenswert hält. Diese können durchaus inkonsistent oder nicht erreichbar sein. Erst wenn der Agent sich auf einen oder mehrere dieser Zustände festlegt, werden sie zu seinen **Zielen** (zu dieser Festlegung oder auch *commitment* siehe den nächsten Abschnitt).
- **Intentions** (Absichten) sind die ausgewählten Ziele, die ein Agent im Moment verfolgt.

Es gibt eine Reihe von Systemen, die diese Architektur benutzen (siehe z.B. (Bratman et al., 1987), (Rao und Georgeff, 1991)). Die bekannteste Vertreter sind aber wohl die Systeme TILEWORLD und MyWorld. Wie ein Testbett im allgemeinen unterstützen sie die Implementierung durch bereitgestellte grundlegende Funktionalität für Multi-Agenten Systeme. TILEWORLD ist eines der bekanntesten solcher Entwicklungssysteme in der VKI. Es erlaubt die einfache Implementierung von Agenten in einer zweidimensionalen Welt. Ihre gemeinsame Aufgabe besteht darin Kacheln einzusammeln und sie auf zufällig auftretende Löcher in der TILEWORLD zu legen. Dabei haben alle Agenten beschränkte Ressourcen, sie können nur beschränkt viele Kacheln tragen und haben eine beschränkte „Lebenszeit" (Pollack und Ringuette, 1990). MyWorld bietet ebenfalls eine Entwicklungsumgebung für BDI-Agenten und legt besonderen Wert darauf, auf den Ansatz des agenten-orientierten Programmierens von Shoham (1993) einzugehen (Wooldridge, 1995).

Ein prominenter Vorwurf gegenüber BDI-Architekturen ist die mangelnde Fähigkeit, während der Planung auf überraschende Ereignisse schnell zu reagieren. Dies war die Motivation zur Entwicklung von Architekturen, die Reaktivität und Deliberation integrieren. Als Beispiel für ein solches hybrides System ist InteRRaP zu nennen, das eine BDI-Architektur implementiert (Müller und Pischel, 1993). InteRRaP Agenten bestehen aus einer Schnittstelle zur Welt, einer Wissensbasis, und einer Kontrolleinheit. Diese Kontrolleinheit setzt sich wiederum aus drei verschiedenen Ebenen zusammen: kooperative Planungsebene, lokale Planungsebene und schließlich verhaltensbasierte Ebene. Letztere Ebene ermöglicht reaktives Verhalten unter Echtzeitbedingung. Die beiden anderen Ebenen sind deliberativ. Die lokale Planungsebene plant die Aktionen, die nur den Agenten selbst betreffen. Die kooperative Planungsebene übernimmt die kooperative Planung und, direkt im Zusammenhang dazu, die Kommunikation mit anderen. Sundermeyer nennt weitere dieser hybriden Systeme (Sundermeyer, 1993): das System IRMA und das für BDI-Architekturen prototypische PRS. IRMA (*Intelligent Resource bounded Machine Architecture*) ist eine absicht-orientierte Architektur, die Intentionen zum einen funktional als Bestandteile von Plänen auffaßt und zum anderen verwendet, um Handlungsoptionen zu filtern (Pollack und Ringuette, 1990). Das PRS (*Procedural Reasoning System*) wurde für einen mobilen Roboter entwickelt. Es handelt sich dabei um einen partiellen, hierarchischen Planer, bei dem Planung und Ausführung miteinander verwoben sind (Georgeff und Ingrand, 1990).

Diese Systeme haben gemeinsam, daß sich die Agenten dort *rational* verhalten. Dies bedeutet, daß sie ihr Verhalten im Hinblick auf eine Nutzenfunktion optimieren. Dies setzt implizit voraus, das ein Agent Ziele verfolgt, deren Erreichen durch die Nutzenfunktion getrieben wird. Rosenschein und Kaebling nennen einen

Agenten rational, wenn er sich optimal im Hinblick auf seine Ziele verhält (Rosenschein und Kaebling, 1986). Neben der Konstruktion von rationalen Agenten (und den etwas exotischeren emotionalen Agenten) hat sich der Begriff der *beschränkten rationalen* Agenten von Simon (1955) auch in der VKI durchgesetzt. Solche Agenten zeichnen sich dadurch aus, daß sie explizit mit Beschränkungen ihrer Ressourcen rechnen. Solche Ressourcen können beispielsweise Rechenzeit, Speichergröße, aber auch Ausgabemedien etc. sein (siehe dazu auch Russell und Wefald (1991)).

2.1.2. Kommunikation und Kooperation

Der Bereich der Multi-Agenten Systeme beschäftigt sich aber nicht nur mit der Konstruktion solcher Agenten, sondern auch mit den Effekten, die sich durch die Verteilung der Kontrolle und der dadurch notwendigen Interaktionen zwischen den Agenten ergeben (von Martial, 1992). Kommunikation und Kooperation in Multi-Agenten Systemen sind zentrale Problemstellungen in der Forschung. Zur Kommunikation von Objekten über Rechner- und Betriebssystemgrenzen hinweg haben sich verschiedene Standards herausgebildet, die auch im großen Rahmen von Wissenschaft und Wirtschaft unterstützt werden.

Einer dieser Standards ist CORBA, der von der OMG (*Object Management Group*) entwickelt wurde. Die OMG wurde 1989 von SUN und anderen Firmen gegründet und hat heute über 700 Mitglieder. Ziel der OMG ist es mit CORBA ein Objektmodell zu entwickeln, das die Portabilität, Wiederverwendbarkeit und Interoperabilität von Softwarekomponenten ermöglicht. Jede Softwarekomponente, also auch ein Agent, wird von der OMG als *Objekt* angesehen. Diese Objekte sollen programmiersprachen- und betriebssystemunabhängig miteinander kommunizieren. Der Entwickler soll darüber hinaus nicht merken, ob die von ihm referenzierten Objekte auf dem lokalen Rechner liegen oder über ein Netzwerk transportiert werden müssen. Mittlerweile ist CORBA als Standardarchitektur von ISO und X/Open anerkannt (Vossen, 1997). Ein weiterer Standard mit derselben Zielsetzung ist DCOM von Microsoft. Um die Kommunikation zwischen Agenten im Sinne einer gemeinsamen Kommunikationssprache auf einer oder mehreren Plattformen zu ermöglichen, wurde KQML entwickelt (Finin et al. 93). Als kommerzielle Alternative hat sich der Standard der *Foundation for Intelligent Physical Agents* (*FIPA*) etabliert. Beide Ansätze beruhen auf der Sprechakttheorie (*speech act theory)* von John Searle (Searle, 1969).

Wenn ein Agent ein genügend genaues Modell des Verhaltens anderer Agenten hat und über entsprechende Sensorinformationen verfügt, so haben Rosenschein et al. gezeigt, ist Kooperation auch ohne Kommunikation möglich. In ihrem spieltheoretischen Ansatz haben sie Agenten die Ergebnismatrizen (ähnlich der Bewertungsfunktion) anderer Agenten und die Daten über deren Verhalten zugänglich gemacht (Genesereth et al., 1984). Die Autoren stellen dann eine Analyse von fast allen möglichen Kombinationen von Werten für die Ergebnismatrizen vor und berechnen die rationalen Entscheidungen für das zwei-Personen-Spiel. Es schränkt jedoch diesen Ansatz ein, daß allen Agenten schon im vorhinein die Punkterträge von allen Agenten bekannt sein müssen.

Im Regelfall wird Kommunikation jedoch als notwendige Grundlage der Kooperation angesehen. Jones definiert Kooperation durch die Abgrenzung zu den Begriffen Koordination und Kollaboration (Jones 90):

- *Kooperation* ist die Zusammenarbeit von Agenten bezüglich einer Aufgabe, wobei sie sich den Profit von dieser Aufgabe teilen.
- *Kollaboration* ist die Zusammenarbeit an einer Aufgabe. Dabei findet ein Aufteilen des Profits aus dieser Aufgabe möglicherweise nicht statt (weil es z.B. gar nicht möglich ist).
- *Koordination* ist das Einwirken auf einzelne Teile, so daß sie in einer vernünftigen Art und Weise ihre Handlungen aufeinander abstimmen.

Castelfranchi et al. verfeinern diese Definition durch eine Unterteilung des Begriffs Kollaboration (Castelfranchi et al., 1997). Ihrer Meinung nach gibt es davon prinzipiell zwei Formen: die *Delegation* und die *Adoption*. Unter Delegation verstehen sie die Abgabe einer Aufgabe (z.B. einen Teilplan) an einen anderen Agenten. Adoption ist die Übernahme eines Ziels von einem anderen Agenten.

Als Realisierungen einer Architektur, die insbesondere die Agenten-kommunikation und –kooperation unterstützt, ist das System MECCA zu nennen. Im System MECCA sind Agenten als bestehend aus Körper, Kopf und Mund modelliert. Der Körper steht für die Fähigkeiten des Agenten, der Mund für die Kooperation. Die Implementierung des Kopfes ist die Modellierung der Agentenarchitektur. Das System stellt Kooperationsprimitive zur Verfügung, die vom Planer wie Aktionen behandelt werden. Sie besitzen also Vor- und Nachbedingungen. Dadurch kann der Planer über die Kommunikation mit anderen Agenten wie über andere Aktionen planen (Lux und Steiner, 1995). Das am DFKI entwickelte *Social Interaction Framework SIF* unterstützt ebenfalls die Kommunikation zwischen Agenten und wurde in Hinblick auf die Trennung der Implementierung des Agenten von der eigentlichen Simulationsmaschine entwickelt. Dadurch kann es in besonderem Maße die Wiederverwendbarkeit von Modulen für neue Szenarien bereitstellen (Schillo et al., 1999a).

Kooperation und der Weg der Kompromißfindung werden traditionell als Verhandlungen modelliert. Für Verhandlungen sind eine Reihe von verschiedenen Protokollen (formale Beschreibungen des Vorgehens) entwickelt worden, die vor Beginn einer Verhandlung verabredet werden. Ziel einer Verhandlung ist es, daß einer oder mehrere Agenten sich (rational) zu einer Handlung entscheiden, deren Ergebnis von anderen genutzt werden kann. Diese Verpflichtung wird auch als *commitment* bezeichnet. Cohen und Levesque haben diesen Begriff präzisiert. *Commitment* ist für sie ein „relativiertes und persistentes Ziel“. Um ein *commitment* zu sein, wird das Ziel als Zustand formuliert (daher *relativiert*), der vom Agenten solange angestrebt wird (daher *persistent*), bis das Ziel entweder nicht mehr erwünscht ist, nicht mehr erfüllt werden kann oder aber erreicht worden ist ((Cohen und Levesque, 1987), (Cohen und Levesque, 1990)).

Kooperation ist ein sich immer noch entwickelndes Gebiet der Verteilten Künstlichen Intelligenz. Bond und Gasser sahen schon 1988 das Problem der Multi-Agenten Systeme darin, daß es bei *Offenen Systemen* dazu kommen kann, daß es weder eine globale Kontrolle noch global konsistentes Wissen, vielleicht noch nicht einmal

eine globale Repräsentation des Systems gibt (Bond und Gasser, 1988). Es hat sich gezeigt, daß sich das Problem der Koordination als Problem der Verhandlung von Aufgaben oder Zielen darstellen läßt. Dadurch ergibt sich die Möglichkeit dieses Problem spieltheoretisch zu behandeln (mehr dazu in Abschnitt 2.4). Im Bereich der VKI-Forschung zur Koordination von Agenten gibt es eine Reihe von Arbeiten auf dem Gebiet der Koalitionsbildung ((Shechory und Kraus, 1993), (Zlotkin und Rosenschein, 1994), (Ketchpel, 1994), (van der Linden und Verbeek, 1985) oder (Kahan und Rapoport, 1984)). Dies sind rationale Agenten. *Beschränkt* rationale Agenten wurden z.B. von Sandholm und Lesser konzipiert (Sandholm und Lesser, 1995). Als Kooperationsmechanismus benutzen sie die Maxime der Agenten, ihren Gewinn zu optimieren und die Suche nach Koalitionen, die dies durch Arbeitsteilung versprechen. Die Autoren teilen die Koalitionsbildung in drei Phasen ein:

1. *Bildung der Koalitionsstruktur*. Die Agenten bilden kooperierende Gruppen. Agenten aus verschiedenen Gruppen kooperieren nicht.
2. *Optimierung der Koalition*. Die einzelnen Aufgaben werden so verteilt, daß das Problem optimal gelöst wird.
3. *Gewinnverteilung*. Die Agenten verteilen am Schluß den erwirtschafteten Gewinn.

Obwohl die einzelnen Vorgänge hier getrennt dargestellt sind, interagieren die Phasen: Ein Agent wird sich erst zu einer Koalition bekennen, wenn ihm klar ist, wieviel er dabei in der letzten Phase bekommt. Dieser Mechanismus zur Kooperation wird in der KI häufig genutzt.

Zeng und Sycara (1996) plädieren für lernende Agenten in Verhandlungen und stellen ihr *Bazaar* System vor. Bazaar unterstützt die Verhandlung von Angeboten mit verschiedenen Attributen, die Modellierung des Verhaltens anderer Agenten und erschließt damit auch die Möglichkeit, offen für Veränderungen im Verhandlungsumfeld zu sein. Sie bauen dabei auf ein sequentielles Verhandlungsprotokoll. Genauer gesagt: Anbieter und Konsument geben nacheinander Angebote ab, solange bis entweder ein Angebot für beide akzeptabel ist, oder aber ein akzeptables Angebot nicht gefunden werden kann. Das Verfahren terminiert.

Die Tauglichkeit von Vertrauen als Kooperationsmechnanismus wurde bisher noch nicht ausreichend untersucht. Mit Ausnahme der Dissertation von Marsh (1994) fehlt bisher ein formales Modell zur Berechenbarkeit von Vertrauenswürdigkeit in Multi-Agenten Systemen. Üblicherweise wird davon ausgegangen, daß ein Agent in einem Verhandlungsprotokoll seine Zusicherung immer einhält, das heißt, daß er sich benevolent verhält. Ein nicht-benevolentes Verhalten wäre zum Beispiel zu betrügen, oder aber für die effiziente Lösung einer gemeinsamen Aufgabe wichtige Informationen absichtlich zu verheimlichen um den eigenen Gewinn zu erhöhen. Rosenschein und Genesereth schreiben zur Benevolenzannahme in Multi-Agenten Systemen:

> *„die implizite Benevolenzannahme erlaubt eine Vereinfachung der Kommunikation und Kooperation zwischen den Agenten, ist aber in der Tat unrealistisch für praxisnahe Anwendungen“*
>
> *(Rosenschein und Genesereth, 1985)*

Das nicht-benevolente Verhalten wird immer dann auftreten, wenn rationale Agenten mit eigenen Interessen dies als subjektiv gewinnbringend ansehen. Als Beispiele für die Notwendigkeit der Modellierung von nicht-benevolenten Systemen können Agenten dienen, die Firmen repräsentieren und durch Zusammenarbeit mit anderen ihre Kosten reduzieren oder den Umsatz erhöhen wollen. In der Praxis zeigt sich z.B. im Speditionsszenario des DFKI, daß Firmen obwohl sie der Zusammenarbeit ja prinzipiell zugestimmt haben, trotzdem nur unrentable Aufträge an Koalitionspartner weitergeben wollen (Bürckert et al., 2000). Als ein weiteres Beispiel kann die Kommunikation sensibler Daten zwischen Agenten in einem Offenen System angeführt werden, denn hierbei ist nicht ausgeschlossen, daß Agenten versuchen, diese Daten zu mißbrauchen, oder aber an andere weitergeben, die sie mißbrauchen.

2.1.3. Maschinelles Lernen

Das grundlegende Probleme aller lernenden Systeme beschreibt Weiß als das *credit-assignment* Problem, d.h. das Problem Lob oder Tadel (*credit* und *blame*) aufgrund von Veränderungen der Systemleistung zu erteilen (Weiß, 1996). Das *credit-assignment* Problem wurde zuerst von Minsky beschrieben (Minsky, 1961). Weiß unterteilt dieses Problem in zwei Teilprobleme:

1. Die Vergabe von Lob und Tadel bezüglich der Leistungsänderung des Systems aufgrund externer Handlungen.
2. Die Vergabe von Lob und Tadel bezüglich der Leistungsänderung des Systems aufgrund interner Prozesse, die zu der externen Handlung geführt haben.

Das erste Teilproblem, welches Weiß *inter-agent credit-assignment* Problem nennt, hält er für besonders schwierig im Bereich der Multi-Agenten Systeme. Da hier viele Agenten potentiell die „falsche“ Handlung ausgeführt haben, oder erst eine Komposition verschiedener Handlungen zu einer Leistungssenkung geführt hat, ist es nicht einfach, den Verursacher auszumachen. Beim zweiten Problem, das er *intra-agent credit-assignment* Problem nennt, sieht er keinen Unterschied ob es sich dabei um einen einzelnen oder mehrere Agenten handelt. Weiß klassifiziert Arten des Lernens nach zwei verschiedenen Kriterien. Das erste Kriterium ist die Lernmethode. Danach unterscheidet er folgende Methoden:

- Auswendiglernen (d.h. dem Agenten werden Wissen und Fähigkeiten direkt eingepflanzt, so daß keine Inferenz oder Transformation mehr notwendig ist).
- Lernen durch Instruktionen und Ratschläge (d.h. Transformation einer neuen Information, etwa einer Instruktion oder eines Ratschlages, in eine für den Agenten ausführbare Operationalisierung).
- Lernen durch Beispiele bzw. *learning by doing* (d.h. die Extraktion und Verfeinerung von Wissen und Fähigkeiten, etwa einem allgemeineren Konzept oder einem standardisierten Bewegungsmuster durch positive und negative Beispiele oder durch Ausprobieren).

- Lernen durch Analogien (d.h. die lösungserhaltende Transformation von Wissen und Fähigkeiten von einem gelösten Problem zu einem ähnlichen, aber noch nicht gelösten Problem).
- Lernen durch Entdeckung (d.h. der Erwerb von neuem Wissen oder neuen Fähigkeiten durch Beobachtungen, dem Ausführen von Experimenten und dem Generieren von Hypothesen).

Für soziale Systeme ist sicherlich noch das Lernen durch Imitation hinzuzufügen. Der Hauptunterschied zwischen diesen Methoden liegt im Aufwand der für das Lernen jeweils aufgebracht werden muß. Das zweites Kriterium nach dem man maschinelles Lernen unterscheidet, ist das Kriterium *Feedback*. Feedback meint die Rückmeldung, die während des Lernprozesses an den Lernenden geschickt wird:

- *Supervised Learning* (d.h. das Feedback spezifiziert die gewünschte Reaktion des Lernenden und das Ziel des Lernens ist, diese Reaktion so gut wie möglich zu antizipieren)
- *Reinforcement learning* (d.h. das Feedback wird nach einer Nutzenfunktion bezüglich der Aktion des Lernenden berechnet und das Ziel des Lernenden ist es, diesen Wert zu maximieren)
- *Unsupervised Learning* (d.h. es gibt kein explizites Feedback und das Ziel des Lernens ist es, auf der Basis von trial-and-error und selbstorganisierenden Prozessen nützliche und erwünschte Aktivitäten zu finden)

Einen guten Überblick über die Arbeiten auf dem Gebiet des maschinellen Lernens im allgemeinen findet sich in Weiß und Sen (1996) oder Mitchell (1997).

2.1.4. Soziale Situiertheit

Der Begriff der „Situiertheit" wurde in der KI der 80er Jahre geprägt. In seiner ursprünglichen Verwendung kann der Begriff „situiert" wohl am Besten mit dem philosophischen „(in eine Situation) geworfen" umschrieben werden ((z.B. (Müller, 1993), (Suchman, 1987), (Agre und Chapman, 1987)). Diese Bedeutung ist auch heute noch in Gebrauch. Situierte Agenten werden von Rao et al. beschrieben als:

> *„…Systeme, die in dynamische Systeme eingebettet sind. Sie nehmen ihre Umwelt über die Zeit wahr und üben Veränderungen durch eigene Handlungen aus. Solche Agenten müssen abwägen, wieviel der ihnen zur Verfügung stehenden Zeit sie in Denken und wieviel sie in Handeln investieren. Außerdem müssen sie die Notwendigkeit der Reaktion auf neue Situationen gegen das Verfolgen von langfristigen Zielen abwägen."*
>
> *(Rao et al., 1992)*

Auf den Begriff *sozial* soll im Abschnitt 2.2 eingegangen werden. Im Zusammenhang mit dem Begriff der Situiertheit greifen wir hier jedoch schon einmal vor. Der Begriff *sozial* taucht in der VKI in verschiedenen Zusammenhängen auf. Unter anderem wurde der Begriff der *sozialen Situiertheit* von Sengers geprägt (Sengers, 1996). Der Begriff wurde von der Arbeit über *glaubwürdige Agenten (believable agents)* inspiriert und sieht Agenten nicht nur als situiert in einem physikalischen sondern auch einem sozialen und kulturellen Umfeld. Die zentrale Zielsetzung des Konzeptes *soziale*

Situiertheit (*situatedness*) ist es, Agenten zu bauen, die interne Zustände ausdrücken können und so zu einem besseren Verständnis ihrer Persönlichkeit und vor allem ihrer Ziele führen.

Dies soll aber nicht durch die Ausgabe eines kryptischen Datenwusts geschehen, sondern durch Mimik, Gestik, Laute etc. In diesem Zusammenhang erweitert sich automatisch auch der Begriff der Handlung (*action)*. Denn von nun an bedeutet eine Handlung möglicherweise die Veränderung des Aussehens des Agenten (lächeln, die „Stirn" runzeln, usw.) und nicht mehr nur die Veränderung des Zustandes der Umwelt wie bei den klassischen Handlungen (fortbewegen, Gegenstände aufheben, usw.). Damit eröffnet sich für Systeme die explizit planen eine neue Dimension im Planungsraum. Es reicht nun nicht mehr geschickt auszuwählen, was zu tun ist, sondern auch wie etwas zu tun ist. Möglicherweise erreicht er sein Ziel viel schneller, wenn er einen anderen Agenten nett um etwas bittet und dabei lächelt, im Vergleich zum Fragen von „gibst Du mir *X*?". Brooks (1986) provozierte durch die These, daß Intelligenz ohne jegliche Repräsentation auskommen kann. Analog dazu zeigen Demiris und Hayes (1997), daß auch soziales Verhalten ohne eine symbolische Repräsentation auskommen kann. Ihr Ansatz beruht auf einer Untersuchung verschiedener Ebenen von Imitation.

Bereits vor dem Auftauchen des Begriffs der sozialen Situiertheit gab es auf Verhalten basierende Programme ((Brooks, 1986), (Maes, 1989), (Blumenberg, 1994), (Agre und Chapman, 1987)). Sengers bemängelt bei diesen die voneinander losgelösten Implementierungen der einzelnen Verhalten, was zur Folge hat, daß die gezeigten Verhalten keinen Bezug zueinander herstellen. Daher wirkt das Gesamtverhalten auf Menschen oft verwirrend. Ihr Verbesserungsvorschlag ist die explizite Verwendung von Verhaltensübergängen (*behaviour transitions*), die zu einem natürlicheren Verhalten führen (eine ausführliche Darstellung der Realisierung befindet sich in (Sengers, 1996)). Sie unterstreicht auch die Notwendigkeit, bei der Modellierung von sozial situierten Agenten die Subjektivität der Entwickler zu beachten. Beim Zuordnen von Verhalten zu internen Zuständen decken sich die Vorstellungen von Experten möglicherweise nicht mit denen der Benutzer, was sehr schnell zum Scheitern eines solchen Systems führen kann.

Auch losgelöst von der sozialen Komponente ist Situiertheit eine wichtige Fragestellung für die Wissenschaft. Der Psychologe Luger (1994) sieht drei wichtige Bereiche, in denen Situiertheit für Agenten eine Rolle spielt. Zunächst kann die Expertise, die ein Agent sich angeeignet hat, kontextabhängig sein. Dies trifft zum Beispiel auf diagnostische Fähigkeiten zu. Im medizinischen Bereich gibt es nur eine Möglichkeit Rückschlüsse aus dem Befund „vergrößerte Leber" zu ziehen, nämlich den Vergleich mit selbst gesammelter Information. Luger erweitert den Begriff der *Bedeutung* wie Tarski ihn sah um eine situative Komponente. Tarski sah Bedeutung als eine Zuordnung von „internen" Repräsentationen zu Objekten, Eigenschaften und Relationen „in der Welt". Luger betrachtet Bedeutung im Zusammenhang mit dem Kontext des Agenten. Bedeutung wird analysiert durch Betrachtung der Typen der Interaktion zwischen Agent und der Welt. Somit kann Bedeutung von Agent zu Agent und von Welt zu Welt verschieden sein. Der letzte Punkt, den er anführt, ist die externe Repräsentation von Gedächtnis, die Agenten benutzen können. Zum Beispiel können Agenten Medien (Papier, *Blackboards* etc.) als Gedächtnisstützen in

der Welt verwenden. Der Agent kann seine Umwelt also vielseitig als Hilfe zur Problemlösung nützen, ist dann aber unter Umständen auch in seinen Fähigkeiten auf sie angewiesen.

2.1.5. Multi-Agenten Systeme: Pro und Kontra

Multi-Agenten Systeme haben Vor- und auch Nachteile. Weiß (1996) zeigt vier prinzipielle Gründe die für die Verwendung von Multi-Agenten Systemen sprechen. Multi-Agenten Systeme sind verteilte Systeme, sie bieten nützliche Eigenschaften wie *Parallelität*, *Robustheit* und *Skalierbarkeit* und sind daher in vielen Bereichen einsetzbar, in denen monolithische Systeme nicht benutzt werden können. Multi-Agenten Systeme sind insbesondere bei Problemen von Vorteil, bei denen es um die Integration mehrerer Quellen von Wissen oder Aktivität, die Auflösung von Interessen- und Zielkonflikten oder das Verarbeiten von großen Datenmengen geht. Nah verwandt hiermit sind die Vorteile des verteilten Planens, die Hertzberg (1989) folgendermaßen beschreibt: Wünschenswert ist *Verteiltheit* in einer Reihe von Szenarien, da es denkbar ist, daß die Information über Teilbereiche an verschieden Stellen vorliegt (es also „Spezialisten" gibt) und es unmöglich oder zu kompliziert ist, alle relevante Information zentral zusammenlaufen zu lassen. Solche Systeme enthalten zwar *Redundanz,* dies kann aber durchaus wünschenswert sein. Technische Systeme dieser Art sind in der Lage Ausfälle von Teilsystemen ohne qualitative Einbußen zu verkraften. Schließlich können mehrere Planer, wenn sie sinnvoll funktionieren, einen Plan zur Lösung eines Problems schneller finden als ein einziger und sie besitzen eine hohe *Zeiteffizienz.* Desweiteren steht das Konzept der Multi-Agenten Systeme im Einklang mit der Einsicht aus Bereichen der Psychologie, der KI und der Soziologie, daß Intelligenz und Interaktion tiefgehend und unausweichlich miteinander verwoben sind. Multi-Agenten Systeme integrieren diesen Gedanken in zweierlei Weise: Zum einen erlaubt die Interaktivität das Verhalten des Gesamtsystems intelligenter erscheinen zu lassen als die Summe seiner Teile. Zum anderen erlaubt ihre Intelligenz es den Agenten, die Interaktion effizienter zu gestalten. Außerdem trägt das Studium der Multi-Agenten Systeme zu unserem Verständnis natürlicher „Multi-Agenten Systeme", wie z.B. Kolonien von Insekten und menschlichen Teams im allgemeinen bei. Insbesondere hilft die Forschung beim Verstehen komplexer sozialer Phänomene wie kollektiver Intelligenz und emergentem Verhalten. Empirische Untersuchungen in diesem Bereich wurden möglich, da die Leistung heutiger Rechner es möglich macht, Multi-Agenten Systeme für solche Studien stabil zu realisieren (Weiß, 1996).

Trotz der oben genannten Vorteile und Eigenschaften sind Multi-Agenten Systeme in einer Reihe von Fällen keine optimalen Problemlöser. Um Problemlösen zu verteilen müssen nach Hertzberg (1989) einige Voraussetzungen erfüllt sein. Er sieht zunächst die Einschränkung der Kommunikation (er betrachtet vor allem das verteilte Planen, was hier stellvertretend für das verteilte Problemlösen angeführt ist). Diese Einschränkung ist sehr wichtig. Zwar können mehrere Agenten vermutlich dann besonders gut zusammenarbeiten, wenn sie unbegrenzt viel kommunizieren können, aber wenn der Aufwand für Kommunikation zu hoch wird, ist die unverteilte Planung einfacher. Außerdem muß ein Problem überhaupt zerlegbar sein. Das Verteilen eines Problems hat nur dann Sinn, wenn das Problem ohne allzu

großen Aufwand in interaktionsarme Teilprobleme zerlegbar ist. Andernfalls ist der Aufwand für die Abstimmung untereinander zu hoch. Eine zentrale Kontrolle gibt es bei Multi-Agenten Systemen, im Gegensatz zum verteilten Problemlösen, nicht. Desweiteren müssen die Agenten ein angemessenes Bild der Fähigkeiten der anderen Problemlöser haben. Wenn die Agenten nichts voneinander wissen ist auch die Verteilung sinnlos, da dann keine Zusammenarbeit zustandekommt. Andererseits ist es auch nicht erstrebenswert, wenn jeder planende Agent von den anderen alles genau weiß und berücksichtigt, was sie tun und wie sie es tun. Der Aufwand würde dadurch nicht verteilt, sondern vervielfacht. Zum Abschluß ist festzustellen, daß es in einem solchen System keine zentrale Kontrolle gibt. Insbesondere dann, wenn man Redundanz erzielen möchte, darf die Architektur des Gesamtsystems keinen der Agenten als Kontrollinstanz bevorzugen. Schwerpunkte dürfen lediglich nach Abstimmung der Agenten untereinander am Problem orientiert vergeben werden. Andernfalls wäre die Kontrollinstanz genau der Flaschenhals, der vermieden werden sollte.

2.2. Soziologie und der Begriff „sozial“

Der Begriff *sozial* ist mittlerweile zu einem Modewort geworden und es gibt sehr viele Ebenen, auf denen er angewendet wird (Dautenhahn et al., 1997). Der Heidelberger Soziologe H. P. Henecka bemüht sich um eine Definition des Wortes *sozial*. Er sieht vier verschiedene Bereiche, aus denen sich die Bedeutung zusammensetzt (Henecka, 1994). Zunächst sieht er hinter dem Begriff *sozial* eine *ethisch-moralische* Haltung, die bis auf den römischen Politiker, Philosoph und Dichter Seneca zurückgeht. Von diesem ist das Zitat überliefert, daß „es sozial sei, ein gutes Werk zu tun“. Diese Bedeutung des Wortes findet sich genauso im christlichen Verständnis, wie in der heutigen Umgangssprache wieder („ein sozialer Mensch“, „einen sozialen Tag haben“). Ein weiteres Bedeutungsfeld ist der *politische* Bereich, genauer gesagt der Bereich, wo Probleme von einzelnen Menschen aufgrund privater ethischer-normativer Verpflichtungen nicht mehr gelöst werden können. Diese Probleme bedürfen einer politischen Lösung und das Wort *sozial* gewinnt eine *öffentlich-politische* Dimension *(z.B. Sozialpolitik, Sozialstaat*). In diesem Zusammenhang steht auch das aus der Theorie von Karl Marx vorkommende, mit *sozial* verwandte Wort *sozialistisch*.

Die für uns im folgenden maßgebende Verwendung ist die *wissenschaftliche*, die nach Henecka gegenüber den bisher aufgezeigten Bedeutungen wesentlich weiter gefaßt ist. Ausgehend von der Grundtatsache, daß der Mensch als *soziales Wesen* von anderen Menschen in hohem Maße abhängig ist, wird als *sozial* hier jedes zwischenmenschliche Verhalten bezeichnet, gleichgültig, ob es sich um „gute“ Taten oder „schlechte“ Formen des Miteinanderumgehens, um moralische Verbundenheiten oder unmoralische Verhaltensakte handelt. Es beinhaltet also nicht nur Werke der Nächstenliebe, sondern auch Akte der Gleichgültigkeit, des Wettbewerbs oder des offenen Konflikts. In deutlichem Gegensatz zum normativen Begriff des *Sozialen,* so Henecka, werde der Begriff im wissenschaftlichen Rahmen *wertneutral* genutzt. Er zitiert Ross: „alle Phänomene, die wir nicht erklären können, ohne dabei den Einfluß des einen Menschen auf den anderen einzubeziehen“. Diese Auffassung von *sozial* dient vorrangig dem Wissensgewinn.

Unter sozialer Intelligenz versteht Aylett die Fähigkeit des Individuums in einem sozialen Umfeld zu „gedeihen“ (Aylett, 1997). Sie kritisiert, daß von vielen WissenschaftlerInnen unter sozialen Fähigkeiten nur die Fähigkeit des kooperativen Problemlösens gesehen wird. Sie argumentiert, daß der Begriff weiter gefaßt ist, da soziale Fähigkeiten auch auf das soziale Umfeld selbst gerichtet sein können, z.B. Status, Hierarchie, usw. Mit anderen Worten: Die sozialen Fähigkeiten können reflexiv auf das Phänomen *sozial* wirken.

Die Definition eines sozialen Umfelds gestaltet sich schwieriger. Van den Berghe definiert Gruppen innerhalb einer Gesellschaft von Individuen durch die Anzahl der Interaktionen: Eine Anzahl von Individuen ist dann eine Gruppe, wenn unter den Gruppenmitgliedern deutlich mehr Interaktionen stattfinden, als zwischen Gruppenmitgliedern und nicht-Gruppenmitgliedern (van den Berghe, 1980). Hierbei gilt es, Anzahl der Interaktionen durch eine meßbare Metrik zu ersetzen.

Schon seit geraumer Zeit bestehen Kontakte zwischen KI-Forschung und Soziologie. So wurde zum Beispiel ein System zur sozialwissenschaftlichen Einstellungsforschung entwickelt, welches einen menschlichen Interviewer ersetzen sollte (Baurmann und Mans, 1984). Die KI ist jedoch nicht für alle Soziologen ein Gebiet von dem sie sich viel erhoffen. Oft heißt es, die Phänomene, die von der Soziologie betrachtet werden, erreichen eine Komplexität die von der KI noch nicht erfaßt werden kann (Malsch et al., 1996). Dabei werden Begriffe wie neue Lebensstile, gesellschaftlicher Wertewandel oder virtuelle Kommunikationsverhältnisse genannt. Collins kommt zum Schluß: „weder die Wissenschaft (*science),* noch Maschinen können soziales Leben modellieren“ (Collins, 1992). Soziologen wie Collins sehen die Zukunft einer Zusammenarbeit darin, die KI bei der Datenauswertung von soziologischen Erhebungen zu benutzen, um Korrelationen und Strukturen in großen Datenmengen sichtbar zu machen. Dies ist sicherlich mit ein Grund dafür, daß Techniken wie Computer-Simulationen noch nie ein ernstzunehmendes Werkzeug in den Gesellschaftswissenschaften waren ((Doran, 1998), für eine Übersicht der wenigen Ausnahmen siehe (Halpin, 1998)).

Malsch et al. (1996) halten diese Einschränkung aber für voreilig. Im Gegenzug zur Benutzung von KI-Techniken in der Soziologie liefert die Soziologie Techniken zur Wissensmodellierung. Damit läßt sich das „bottleneck“ der Wissensakquisition zwar nicht grundsätzlich überwinden, argumentieren Malsch und seine Kollegen, aber mit sozialwissenschaftlichen Methoden könne das Wissen von Fachexperten zuverlässiger erhoben, treffender strukturiert und genauer formalisiert werden. Die Autoren erklären dies dadurch, daß ihrer Meinung nach Wissen sozial konstruiert ist, sich nur aus seiner gesellschaftlichen Erzeugung und Verwendung heraus verstehen läßt und ohne Berücksichtigung von sozialen Normen, kulturellen Werten und „ökonomischen Interessen nicht angemessen modelliert“ werden kann. Morik geht in dieser Auffassung von Wissen noch einen Schritt weiter: „Wissen hat keine endgültige Gestalt, sondern wird laufend ergänzt, umgearbeitet und weiterentwickelt“ (Morik, 1989).

Nach Florian gibt es eine Entwicklung in den beiden Wissenschaften, die der Zusammenarbeit eine neue Komponente geben wird (Florian, 1996). Seiner Meinung nach steht eine "Hochzeit" zwischen KI und virtueller Realität bevor, die die Wissenschaft insbesondere vor das Problem stellen wird, große

Kommunikationsnetze gesellschaftsanalog zu konstruieren und völlig neuartige, heute noch gar nicht absehbare soziotechnische Anwendungen zu erfinden. Zwar glauben nur wenige Soziologen an die Formalisierbarkeit soziologischer Erklärungen, aber es muß sich erst noch herausstellen, inwiefern die Soziologie eine beobachtende bzw. teilnehmende Rolle bei dieser „Hochzeit" spielt. Die Autoren sind nämlich der Meinung, daß wer angemessene Aussagen über das Soziale treffen will, unabhängig vom Abstraktsionsniveau, an der Soziologie nicht vorbeikommt. Dies ist insofern eine Kritik am Mainstream der VKI Forschung, als dieser bis heute überwiegend in der rationalistischen Perspektive befangen ist. Diese ist ungeeignet um soziale Welten als emergente Phänomene zu beschreiben. Bachmann formuliert seine Kritik an der VKI noch schärfer. Er stellt fest, daß die VKI eigene Laientheorie entwickle und eher heimlich versuche, Anregungen aus den („ihr unverständlichen") Texten der Soziologie herauszulesen (Bachmann, 1998).

In den USA bildet sich eine engere Zusammenarbeit zwischen KI-Forschung und Soziologie. Gasser formulierte, daß ohne soziologische Fundierung der KI kein substantieller Fortschritt auf dem Gebiet von Multi-Agenten Systemen möglich ist (Gasser, 1991). Von Martial sieht zumindest die Chance, daß VKI Systeme zur Validierung von Theorien aus Soziologie, Management- und Organisationstheorie ähnliches leisten können, was KI Systeme bei der Bestätigung von Problemlösungsmodellen aus Linguistik, Psychologie und Philosophie bereits geleistet haben (von Martial, 1992). Malsch et al. (1996) warnen die Soziologie davor, sich dem zu verschließen und schließen sich von Martial an. Aufgrund des Präzedenzfalles der Geistmetapher, in der sie die Ursache für eine Paradigmenrevolution in der Informatik sehen, die zu weitreichenden technischen Innovationen geführt hat, schlagen sie die Forschungsrichtung der *Sozionik* vor. Nun stellen sie die Frage, ob die sozionische Forschung nicht einen bahnbrechenden Innovationsbeitrag leisten kann, der ohne die Inspirationskraft der Gesellschaftsmetapher nicht würde gemacht werden können.

2.3. Vertrauen als Forschungsobjekt

Die Literatur zum Thema Vertrauen ist schier unüberschaubar. Eine abschließende Betrachtung der existierenden Veröffentlichungen sprengt bei weitem den Rahmen einer Doktorarbeit (Platzköster, 1990). Dies liegt zum einen am Umfang der involvierten Disziplinen. Soziologie, Psychologie, insbesondere, aber nicht nur, die Sozialpsychologie, die Philosophie und die Betriebswirtschaftslehre (vor allem Marketing) betrachten Vertrauen aus ihren verschiedenen Perspektiven. Trotz dieses großen Interesses bezeichnen eine Reihe von Autoren den Stand der Vertrauensforschung als unbefriedigend (z.B. (Seligmann, 1997) für die Soziologie, (Koller, 1990) für die Sozialpsychologie, (Dasgupta, 1990) für die Ökonomie). Es ist den Wissenschaften nämlich nicht gelungen, eine befriedigende Definition zu finden. Manche Autoren gehen sogar so weit, zu behaupten, daß der Begriff so komplex ist, daß eine adäquate Definition nicht realisierbar ist ((Brückerhoff, 1982), (Narowski, 1974)). Hinzu kommt, nach dem Soziologen Luhmann, daß eine genaue Untersuchung von Vertrauen noch nie ein Thema des „Mainstreams" der Soziologie

war. Weder haben die Klassiker der Soziologie, noch moderne Autoren diesen Begriff in einem theoretischen Kontext benutzt (Luhmann, 1990).

Dieser Abschnitt gibt einen Überblick über den Stand der Vertrauensforschung. Zunächst zeigen wir, wie Vertrauen in anderen Wissenschaften, speziell der Soziologie und Psychologie, gesehen wird. Danach gehen wir auf die vorhandenen Modelle in der VKI ein. Mit einer Zusammenfassung der wichtigsten Begriffe im Zusammenhang mit Vertrauen runden wir diesen Abschnitt ab.

2.3.1. Eine interdisziplinäre Übersicht

Die Definition von Vertrauen, so argumentiert Marsh (1994), ist in zweierlei Hinsicht problematisch: Zum einen sei jeder Mensch ein „Experte" in Sachen Vertrauen, zumindest was das Vertrauen angeht, daß er oder sie selbst anderen entgegenbringt. Den zweiten Grund den er für die vielen existierenden Definitionen anführt ist der, daß es viele verschiedene Arten von Vertrauen gibt. Platzköster (1990) sieht ein weiteres Problem in der Vermengung verschiedener Konzepte mit Vertrauen. Er beobachtet, daß Begriffe wie Kooperation, Zutrauen (als *confidence*), Glaubwürdigkeit oder Verläßlichkeit oft nicht ausreichend in ein klar definiertes Verhältnis zu Vertrauen gestellt werden und die Analyse dadurch erschwert wird.

Die folgende Aufzählung ist eine Zusammenstellung der verschiedenen Arten von Vertrauen von Marsh, die er von dem Psychologen Deutsch (1973) übernommen hat und die von einem psychologischen Standpunkt aus formuliert ist (siehe auch (Golembiewski und McConkie, 1975)):

Vertrauen als Verzweiflung. Falls die negativen Konsequenzen des nicht-Vertrauens so groß oder so sicher sind, dann baut die Entscheidung aus Verzweiflung auf Vertrauen auf. (Das geringere Übel wählen.)

Vertrauen als Konformität. In vielen Situationen wird Vertrauen erwartet und nicht-konformes Verhalten wird schwer geahndet. So kann es vorkommen, das eine Person einer anderen Geld leiht, obwohl sie recht sicher weiß, daß sie dieses Geld nicht wiederbekommen wird, ganz einfach weil es zu diesem Verhalten keine sozial akzeptierte Alternative gibt.

Vertrauen als Unschuld. Eine Entscheidung wird aufgrund mangelnder Kenntnis der Gefahren getroffen. Diese Unschuld kann zurückgeführt werden auf Mangel an Information, kognitiver Unreife, kognitivem Defekt (pathologisches Vertrauen).

Vertrauen als Impulsivität. Ein Verhalten, daß in unangemessener Art und Weise von Erwartungen an die Zukunft gesteuert wird. Kommt z.B. im Gefangenen-dilemma (siehe Abschnitt 2.4) vor, wenn davon auszugehen ist, daß ein bestimmtes Verhalten in Zukunft Sanktionen nach sich zieht.

Vertrauen als (sozialer) Wert. Wenn Vertrauen eine große Rolle für eine Gesellschaft spielt und der Erhalt der Gesellschaft Ziel ist, wird Vertrauen zum Zweck.

Vertrauen als Masochismus. Menschen wollen Erwartungen bestätigt sehen. Daraus resultiert das Verhalten anderen zu mißtrauen, um die eigenen schlechten Erwartungen über sie erfüllt zu sehen ((Rempel und Holmes, 1986), (Boon und Holmes, 1991)).

Vertrauen als Glaube. An eine Entscheidung zu „glauben" entschärft die Schuld an negativen Konsequenzen einer Entscheidung.

Vertrauen als Glücksspiel oder „Risiko eingehen". Dabei spielt es weniger eine Rolle, sich auf etwas zu „verlassen", als Wahrscheinlichkeiten und erwartete Gewinne und Verluste zu kalkulieren

Ur-Vertrauen. Vertrauen darin, daß das eintreten wird, was gewünscht ist und nicht was befürchtet wird.

In der Modellierung, die mit dieser Arbeit geleistet werden soll, wird vor allem der Punkt „Vertrauen als Glücksspiel" eine Rolle spielen, da er derjenige ist, der am ehesten ohne metaphysische Konzepte wie Glauben und Werte auskommt. Außerdem ist die folgende Definition von Vertrauen nach Deutsch wichtig. Laut Marsh ist sie eine der am weitesten akzeptierten.

Definition 1: Vertrauen nach Deutsch (1962)

> Befindet sich ein Individuum in einer Situation, in der ein weiterer Schritt die folgenden Bedingungen erfüllt:
>
> a) der nächste Schritt hat ein möglicherweise gewinnbringendes und ein möglicherweise verlustbringendes Ergebnis
>
> b) dieses Ergebnis hängt von einem zweiten Individuum ab
>
> c) es ist davon auszugehen, daß ohne das zweite Individuum die verlustbringende Variante eher eintreten wird
>
> und entscheidet sich das Individuum trotzdem diesen Pfad weiter zu beschreiten, dann kann man davon ausgehen, daß das Individuum Vertrauen hat.

Diese Definition ist stark von der Psychologie und vom Gefangenendilemma inspiriert. Insbesondere hebt sie hervor, daß Vertrauen dazu dient, einen *Nutzen* zu haben. Andere Arbeiten greifen diesen Aspekt auf und beschäftigen sich mit dem Phänomen Vertrauen mehr aus ökonomischen Sicht:

Vertrauen als nützlicher Mechanismus. Um Transaktionen jeglicher Art vorzunehmen, ist Vertrauen nützlicher als andere Formen des Verhaltens zur Herstellung von Sicherheit. Für die Ökonomie ist Vertrauen das zentrale Element aller Transaktionen (Dasgupta, 1990). Golembiewski und McConkie sehen Vertrauen als notwendig für jegliche Kooperation (Golembiewski und McConkie, 1975).

Trotz all dieser unterschiedlichen Auffassungen sieht Koller drei Definitionselemente, die sich immer wieder finden:

1. Eine Vertrauensperson wird positiv und eine Person des Mißtrauens wird negativ bewertet.
2. Diese positive Bewertung ist mit der Erwartung verknüpft, daß sich die Vertrauensperson wohlwollend verhalten wird.
3. Dies führt zu einem dritten Element: Eine vertrauensrelevante Situation beinhaltet *Risiko*. D.h. der Interaktionspartner muß sich nicht notwendigerweise wohlwollend verhalten. Er hat auch die Möglichkeit eine Verhaltensalternative zu wählen, die für das vertrauende Individuum mit negativen Konsequenzen

verbunden ist. Das vertrauende Individuum steht bis zu einem gewissen Grad unter der Kontrolle seines Interaktionspartners.

Dies läßt die Möglichkeit einer integrierende Definition wieder realistischer erscheinen. Koller schlägt die folgende Definition vor:

Definition 2: Vertrauen nach Koller (1990)

Vertrauen ist die Erwartung, daß ein Interaktionspartner wohlwollendes Verhalten zeigen wird, obwohl dieser die Möglichkeit hat, andere, nicht-wohlwollende Verhaltensweisen zu wählen.

In dieser Definition taucht der Begriff der *Erwartung* auf, den auch andere Autoren in engem Zusammenhang mit Vertrauen sehen (siehe auch dazu Luhmann im nächsten Abschnitt). Platzköster sieht Erwartungen als die Voraussetzung für Vertrauen. Denn je nach Erwartungen in einen anderen Akteur stellt sich Vertrauen oder Mißtrauen ein (Platzköster, 1990). Rotter, der neben Deutsch wohl zu den wichtigsten Autoren zum Thema Vertrauen in der Psychologie gehört, sieht diesen Zusammenhang ähnlich. Seiner Ansicht nach ist Vertrauen eine geplante Disposition, gebildet aus Erfahrungen mit der Verläßlichkeit. Dies bedeutet, daß ein generalisiertes Vertrauen Verhalten, Erwartungen und Einstellungen in mehrdeutigen und neuen Situationen beeinflußt. Die Zuverlässigkeit von Erwartungsbestätigungen ist also Ursache von Vertrauen (Rotter, 1971). Diese Ansicht ist in anderen Bereichen der Psychologie ebenfalls vorhanden. Sie wird außerdem von dem Soziologen Luhmann vertreten. Er hält komplexe und längerfristige Erwartungen für lebensnotwendig, um überhaupt handeln zu können. Außerdem ist Vertrauen gerade in komplexen Situationen oft der einzige Ausweg, um zu einer Entscheidung zu kommen:

Vertrauen als Mittel zur Komplexitätsreduktion. Es existieren Situationen (Interaktionen) in denen nichts über den Interaktionspartner bekannt ist. In solchen Situationen verhindert ein Überdenken aller möglichen Folgen eine Interaktion. Nur wer die Folgenabschätzung ausläßt (also *vertraut*) ist hier überhaupt fähig zu handeln (Luhmann, 1973).

Luhmann (1979) beschreibt *Risiko* als wichtige Voraussetzung von Vertrauen. Diese Auffassung teilen auch andere Autoren wie Gambetta (1990b) und Marsh (1994). Für Luhmann ist Vertrauen notwendigerweise risikobehaftet, denn nur durch die Unvollständigkeit der aktuell verfügbaren Information müssen Hypothesen über das zukünftige Handeln anderer gemacht werden. Erst bei Unsicherheit handelt es sich nicht mehr um eine kühle Vorausberechnung sondern um Vertrauen. Luhmann geht auch auf ein anderes wichtiges Konzept ein, das die Komplexität einer Entscheidung reduziert, nämlich *Macht*. Verfügt eine Instanz über die Macht Sanktionen auszuüben, wie beispielsweise ein Rechtssystem, so reduziert dies die Komplexität der Welt. Bestimmte Hypothesen können dadurch ausgeschlossen werden oder werden zumindest sehr unwahrscheinlich. Das Vorhandensein von Macht reduziert also die Komplexität der Welt, indem es sie berechenbarer macht. Dadurch wird also weniger Vertrauen in den Interaktionspartner notwendig. Es wird ersetzt durch ein Vertrauen in diese Instanz der Macht. Darin sieht er den direkten Zusammenhang zwischen Vertrauen und Macht.

Direkt mit Vertrauen verknüpft ist auch der Begriff der *Kommunikation*. Insbesondere in Studien zu Organisationsvertrauen zeigt sich, daß nicht nur

Vertrauen sich begünstigend auf die Kommunikation auswirkt, sondern daß auch umgekehrt Kommunikation selbst Vertrauen schaffen kann. Bei der Betrachtung von Vertrauen und Kommunikation wird das Vertrauen in einen *Kommunikator* unterschieden von dem Vertrauen in das kommunizierte. Dabei steht *Glaubwürdigkeit* in so engem Zusammenhang mit Vertrauen, daß sich bisher keine einheitliche Trennung herausgebildet hat (Platzköster, 1990).

An dieser Stelle soll noch auf den Begriff *Mißtrauen* eingegangen werden. Nach Zand ist Mißtrauen dokumentiert durch Verheimlichung bzw. Verzerrung von Information. Mißtrauen bewirkt, daß keine Informationen mitgeteilt werden, die die eigene Verwundbarkeit (Machtverlust) erhöhen könnten. Dabei wird versucht Einfluß- bzw. Machtverlust an andere, die nicht mehr vertrauenswürdig erscheinen, zu vermeiden (Zand, 1977). Gambetta weist darauf hin, daß Vetrauen und Mißtrauen sich bezüglich ihres Maßes zueinander symmetrisch verhalten. Eine Verringerung des Vertrauens bedeutet also gleichzeitig ein Anwachsen des Mißtrauens. Nicht symmetrisch hingegen ist das Ausmaß von Veränderungen im Vertrauen. Während es sehr leicht ist Beobachtungen zu machen, die Mißtrauen rechtfertigen, ist es aber sehr schwierig Verhalten zu bestimmen, aus dem sich zwingend Vertrauen ableiten läßt. Damit einher geht, daß Vertrauen sehr schwer aufzubauen ist, aber sehr leicht zu zerstören ist (Gambetta, 1990b).

Viele der hier vorgestellten Definitionen widersprechen sich oder haben nur geringfügige Gemeinsamkeiten. Außerdem sind sie nur sehr schwer mathematisch formalisierbar. Da dies für unsere Arbeit aber eine notwendige Voraussetzung ist, wollen wir im nächsten Abschnitt besonders auf zwei Vertrauensansätze eingehen, die in der KI-Forschung entwickelt wurden, bzw. gerade entwickelt werden.

2.3.2. Vertrauen in Multi-Agenten Systemen

In der Forschung der Multi-Agenten Systeme haben sich drei Bereiche bezüglich der Vertrauensforschung herausgebildet, die unterschiedliche Anforderungen an ein Modell von Vertrauen haben. Der erste Bereich umfaßt das Vertrauen von Menschen in künstliche Agenten. Dieser Bereich läßt sich grob der Mensch-Maschine-Interaktion (*Human Computer Interaction, HCI*) zuordnen. Hier wird untersucht, welche Maßnahmen notwendig sind, damit ein Software System auf einen Benutzer vertrauenswürdig wirkt. Dies hat keineswegs nur kosmetische oder kommerzielle Hintergründe. Muir (1987) konnte in ihren Studien nachweisen, daß Benutzer, die einem Diagnosesystem nicht vertrauten, ihm solange manipulierte Informationen eingaben, bis es die von ihnen bevorzugte, aber falsche Lösung produzierte (Muir, 1987). Ohne die Berücksichtigung dieses Verhaltens ist die Anwendbarkeit von Expertensystemen, die durch Menschen bedient werden, in kritischen Situationen erheblich in Frage gestellt. Der zweite Bereich befaßt sich damit, wie die Interaktion per Computer zwischen Menschen so unterstützt werden kann, daß die Bildung von Vertrauen nicht durch die Interaktion per Computer behindert wird (*Computer Supported Collaborative Work, CSCW*, bzw. *Computer Mediated Interaction, CMI*).

Der dritte (und in dieser Arbeit behandelte) Bereich schließlich beschäftigt sich damit, wie Vertrauen zwischen Softwaresystemen modelliert und angewendet werden

kann. Stellen wir uns beispielsweise zwei Agenten vor, die die Möglichkeit haben, eine Aufgabe gemeinsam oder allein zu lösen. Bei einer Kooperation soll es von Nachteil für einen Agenten sein, wenn der andere ihn trotz Zusicherung nicht hilft. Die Frage ist nun: Soll ein Agent nun auf die Kooperation eingehen, wenn der andere ihm Kooperation anbietet? Wenn diese Entscheidung von mehr als nur dem Zufall abhängen soll, muß es Kriterien geben, nach denen sie gefällt wird. Welches sind diese Kriterien und wie wichtig sind sie jeweils für die Gesamtentscheidung? Wie ergeben sie gemeinsam die Entscheidung, ob der Agent sich auf das Kooperationsangebot einlassen soll? Die Antworten auf diese Fragen aus den Gesellschaftswissenschaften lassen wie im letzten Abschnitt gezeigt an Formalisierbarkeit zu wünschen übrig. Der Informatiker Marsh dagegen hat einen Formalismus angegeben, von dem er hofft, daß er als Grundlage für die Diskussion über die Funktionsweise von Vertrauen dienen kann (Marsh, 1994). Dieser Formalismus stützt sich auf zu beobachtende Fakten und berechnet daraus die Vertrauenswürdigkeit eines Agenten. Es handelt sich dabei um einen probabilistischen Ansatz.

Für eine strukturierte Zusammenarbeit von Agenten müssen nach Marsh prinzipiell drei Fragen beantwortet werden. Mit wem wird kooperiert? Wie weit geht die Kooperation? Und unter welchen Umständen wird kooperiert? Um algorithmisch entscheiden zu können ob eine Kooperation gewinnbringend ist und in dieser Frage das Konzept Vertrauen einbringen zu können, macht Marsh folgende Annahmen über die Situation, in der ein Agent *X* eine solche Entscheidung treffen muß:

- *X* hat die Möglichkeit zu wählen: *X* muß nicht kooperieren.
- Es existiert ein Agent *Y* mit dem *X* kooperieren kann.
- *X* kennt *Y*.[1]
- *X* ist nicht in der Schuld von *Y*. Damit wird vermieden, daß ein bloßes revanchieren stattfindet (siehe Definitionen von Vertrauen in Abschnitt 2.3.1).
- *X* hat ein Wissen über die Situation, d.h. *X* erkennt Parallelen zu bereits durchlebten Situationen.

Dabei sind die beiden letzten Annahmen im direkten Gegensatz zu den Theorien von Luhmann und Deutsch. Marsh benötigt diese Annahmen jedoch um sein mehrschichtiges Modell von Vertrauen zu konstruieren. Für Marsh gibt es drei Ebenen von Vertrauen. Zunächst kann ein Agent Vertrauen in „Vertrauen an sich" haben. Dies ist also die Einschätzung des Agenten, inwiefern das Vertrauen in andere für ihn sinnvoll ist (*basic trust*). Desweiteren hat ein Agent die Möglichkeit anderen Agenten zu vertrauen. Er stellt also Vermutungen an, wie sehr er einem Agenten im allgemeinen vertrauen kann, unabhängig davon, ob er Vertrauen an sich für rational hält oder nicht (*general trust*). Schließlich gibt es die Ebene des situationsbezogenen Vertrauens. Darunter versteht Marsh das Vertrauen in einen Agenten in einer bestimmten Situation (*situational trust*). Theoretisch kann dies unabhängig von dem

[1] Wie später dargelegt, soll diese Annahme von Marsh im vorliegenden Ansatz nicht mehr notwendig sein.

Vertrauen in den Agenten an sich sein. Nach Marsh berechnet sich das Vertrauen eines Agenten X in einen Agenten Y in einer Situation s folgendermaßen.

$$Vertrauen_X(Y,s) = Nutzen_X(s) \times Wichtigkeit_X(s) \times Vertrauen_X(Y). \quad (1)$$

Das Vertrauen von X in Y in Situation s ergibt sich nach Marsh aus dem Produkt des Nutzens der Situation für X, der Wichtigkeit der Situation s für X und dem generellen Vertrauen das X in Y hat. Dieses generelle Vertrauen setzt sich zusammen aus allen Erfahrungen die X mit Y bisher gemacht hat. Ob X nun Y in der Situation s vertraut, hängt davon ab, ob $Vertrauen_X(Y,s)$ größer als ein Schwellenwert[1] ist. Dieser berechnet sich wie folgt:

$$Schwellenwert = \frac{wahrgenommenesRisiko_X(s)}{wahrgenommeneKompetenz_X(Y,s) + \overline{Vertrauen_X(Y)}} \times Wichtigkeit_X(s) \quad (2)$$

Marsh benutzt das Intervall [-1;1[für seine Werte von Vertrauen. Bei den Korrekturtabellen, die er für manuelle Vorzeichenänderungen in seinen Formeln angibt, übersieht er, daß es auf das üblichere Intervall [0;1[isomorph abgebildet werden kann, in dem auf den Wert x die Umrechnung $\frac{x+1}{2}$ angewendet wird.

Marsh wendet diesen Formalismus auf die zwei Domänen "Möbelpacker" und "PlayGround" an. In der ersten Domäne zeigt er durch Beispielrechnungen, wie einzelne "Möbelpacker" berechnen, wem sie vertrauen ihnen beim Transport eines Möbelstücks, zu helfen. In der zweiten Domäne können Agenten sich frei durch einen Raum bewegen und sich dort einen Partner dem sie vertrauen, für eine Interaktion auswählen. Als Interaktion dient das Gefangenendilemma, die Situationen sind gekennzeichnet durch verschiedene Ergebnismatrizen (siehe Abschnitt 2.4.2). Dort untersucht Marsh inwiefern die Anwendung obiger Formel dazu führt, daß die richtigen Agenten miteinander kooperieren. Es zeigt sich in der Tat, daß Agenten häufig miteinander interagieren, die sich gegenseitig für vertrauenswürdig halten. Da er jedes Experiment jedoch nur einmal ausführt, der Ausgang des Experimentes von der Anzahl der durchgeführten Bewegungen im Raum abhängt und er die Agenten zufällig im Raum verteilt, ist die statistische Sicherheit dieser Aussage nicht gewährleistet.[2]

Marshs Ansatz wird auch von Castelfranchi et al. angegriffen. Ihre Hauptkritik ist die „Flachheit" von Marshs Modell und die Beschränktheit auf ein größtenteils quantitatives Modell. Dadurch entfällt ein großer Teil seiner Arbeit auf die Wahl von Koeffizienten und Vorzeichen in den Berechnungen, die teilweise willkürlich und wenig elegant erscheinen. Castelfranchi et al. dagegen bieten eine reichhaltigere Theorie. Sie besteht aus zwei Teilen: Einer qualitativen Theorie (*was ist Vertrauen*), die sie „kognitive Anatomie" nennen, und einer quantitativen Theorie (*wie berechnet man Vertrauen*).

---

[1] Eigenartigerweise läßt sich aus dieser Gleichung aber die Wichtigkeit der Situation kürzen. Auf diese Tatsache geht Marsh nicht ein.

[2] Weitere Kritik zu diesem Ansatz findet sich in Abschnitt 3.1.

Qualitative Theorie. Nach Castelfranchi et al. (1997) ist Vertrauen ein *mentaler Zustand*. Im Unterschied dazu bezeichnen sie *Delegation* als eine Handlung. Bei einer Delegation werden Ziele oder Handlungen von einem Vertrauenden an einen anderen Akteur abgegeben. Während Delegation ein Prädikat ist (entweder der Vertrauende delegiert oder nicht), drückt Vertrauen die Wahrscheinlichkeit aus, mit der der Vertrauende davon ausgeht, daß die delegierte Handlung ausgeführt wird bzw. das delegierte Ziel von dem anderen Akteur verfolgt wird. Dabei sollte ein wichtiger Bezug dieser Auffassung von Vertrauen zu Multi-Agenten Systemen erwähnt werden: Dadurch, daß Agenten etwas für jemand anderen tun (z.B. für ihren Benutzer) ist Delegation per Definition in MAS enthalten. Jegliche Kooperation unter Agenten enthält wiederum implizit Delegation von Aufgaben und Zielen. Daher kommen die Autoren zu dem Schluß, daß Vertrauen implizit in Multi-Agenten Systemen enthalten ist. Außerdem fassen sie den Begriff der Delegation sehr weit. Sie sprechen von *starker* und *schwacher Delegation*. In der schwachen Delegation gibt es kein gegenseitiges Verabreden. Sie bezeichnen auch das Vertrauen auf ein Objekt oder Werkzeug als schwache Delegation einer Aufgabe. Als Beispiel führen sie einen Jäger an, der mit Pfeil und Bogen einen Vogel erlegen will. Um das Tier zu treffen, zielt der Jäger auf einen Punkt, der auf der zukünftigen Flugbahn des Vogels liegt. Obwohl sich der Vogel dessen nicht bewußt ist, delegiert der Jäger die Aufgabe zu diesem Punkt zu fliegen an den Vogel. Erst die starke Delegation ist umgangssprachliche Definition, die eine Absprache (etwa eine Verhandlung) voraussetzt.

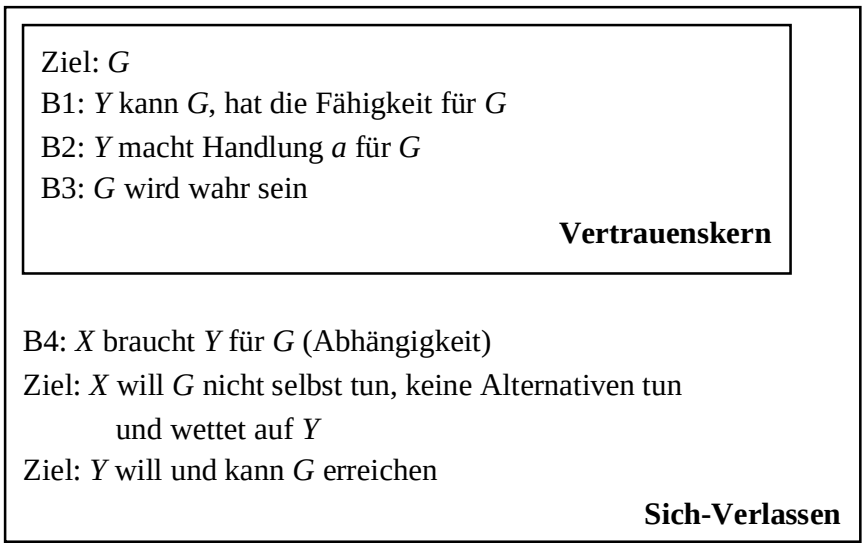

Abbildung 2: Kognitive Anatomie des Vertrauens nach Castelfranchi und Falcone

Wie schon gesagt, Vertrauen ist ihrer Definition nach die Wahrscheinlichkeit, mit der die delegierte Handlung durchgeführt wird. Das qualitative Modell eines Agenten *X* von Vertrauen in einen anderen kognitiven Agenten *Y* bezüglich der Erfüllung eines Zieles *G* besteht aus den folgenden Zielen und *beliefs* (siehe Abbildung 2). Der Vertrauenskern besteht aus den abgebildeten drei *beliefs* (B1: Kompetenz, B2: Intention, B3: Erwartung). Die Erweiterung des Kerns, nämlich das sich-auf-jemanden-verlassen, umfaßt zusätzlich die *Entscheidung* des sich-Verlassens. Sie besteht außerdem aus dem *belief* von *X*, daß *X* von *Y abhängig* ist, daß *X* das Ziel hat,

G nicht selbst tun zu wollen und das Ziel zu erreichen, daß *Y* das Ziel *G* erreichen will und auch erreichen kann. Vertrauen (als Verb) bedeutet dann, daß ein Agent diesen Vertrauenskern besitzt und die Entscheidung zur Delegation getroffen hat.

Definition 3: Vertrauen nach Castelfranchi et al.: Qualitative Definition
Es seien *K* und *B* zwei Prädikate für *Wissen* und *belief*. Außerdem seien die folgenden Prädikate gegeben: *Goal* (Ziel), *PracPoss* (Fähigkeit, Möglichkeit), *Prefer* (vorziehen), *Done* (schon getan), *Intend* (beabsichtigt) und *Persist* (Beständigkeit der Absicht). Dann ist das Vertrauen von *X* in *Y* bezüglich einer Aufgabe τ in prädikatenlogischer Syntax folgendermaßen definiert:

$$\begin{aligned} Trust(X,Y) = K_X Goal_X G \mid & B_X \mathrm{Prac}Poss_Y(\alpha,G) \mid \\ & B_X \mathrm{Prefer}_X(Done_Y(\alpha,G) \mid Done_X(\alpha,G)) \mid \\ & [(B_X(Intend_Y(\alpha,G) \mid Persist_Y(\alpha,G)) \int \\ & (B_X(Intend_Y(\alpha,G) \mid Persist_Y(\alpha,G))] \end{aligned} \qquad (3)$$

Quantitative Theorie. Die Autoren formulieren ihre Theorie auch als Produkt von Wahrscheinlichkeiten.

Definition 4: Vertrauen nach Castelfranchi et al.: Quantitative Definition
Es seien die folgenden Prädikate definiert:

$Opp_Y(\alpha, G)$	*Y* hat die Gelegenheit für eine Aktion α bzgl. eines Zieles G.
$Ability_Y(\alpha, G)$	Fähigkeit einer Aktion α bzgl. eines Zieles G.
$Intend_Y(\alpha, G)$	Absicht eine Aktion α bzgl. eines Zieles G zu machen[1].
$Persist_Y(\alpha, G)$	Wahrscheinlichkeit, daß die Absicht eine Aktion α bzgl. eines Zieles G durchzuführen, erhalten bleibt.

Gegeben sei weiterhin eine Funktion *DoC* (Degree of Credibility - Maß der Wahrscheinlichkeit, daß ein *belief* korrekt ist) die für ein Prädikat angibt, wie wahrscheinlich es ist, daß dieses Prädikat erfüllt ist. Dann berechnet sich das Maß des Vertrauens von *X* darin, daß *Y* die Aufgabe τ ausführen wird folgendermaßen:

$$\begin{aligned} DoT_{XY\tau} = DoC_X(Opp_Y(\alpha,G)) \leftarrow & DoC_X(Ability_Y(\alpha)) \\ \leftarrow & DoC_X(Intend_Y(\alpha,G)) \leftarrow DoC_X(Persist_Y(\alpha,G)) \end{aligned} \qquad (4)$$

Dabei kommt dem Glaube an das Zutreffen von *Intend* die Funktion des *situational trust* von Marsh zu und man könnte argumentieren, daß *Persist* den *global trust* von Marsh darstellt. Ob ein Agent delegiert (ob er vertraut), hängt von dem Verhältnis des *DoT* zu dem erwarteten Nutzen der Delegation ab. Es seien:

---

[1] Das Maß des Glaubens, daß dieses Prädikat erfüllt ist, von dem die Autoren nicht angeben, wie es berechnet wird, soll in dieser Arbeit mathematisch bestimmt werden.

$U(X)_{P^+}$	Der Nutzen bei nicht-Delegation (*Performance*).
$U(X)_{P^-}$	Der Schaden bei nicht-Delegation.
$U(X)_{D^+}$	Der Nutzen bei Delegation.
$U(X)_{D^-}$	Der Schaden bei Delegation.

Gewichtet man die Werte der Nutzenfunktion nach der Bayes'schen Formel, dann empfiehlt sich Delegation falls:

$$DoT_{XY\tau}U(X)_{D^+} + (1 - DoT_{XY\tau})U(X)_{D^-} \quad \textit{größer als} \quad DoT_{XX\tau}U(X)_{P^+} + (1 - DoT_{XX\tau})U(X)_{P^-} \tag{5}$$

Dabei steht $DoT_{XX\tau}$ für das Vertrauen des Agenten in sich selbst. Dies ergibt ein *kognitives* und ein *mathematisches* Model von Vertrauen in die Delegation von Handlungen oder Zielen. Dieses ist sowohl implementierbar, als auch psychologisch fundiert ((Castelfranchi et al., 1997), (Castelfranchi und Falcone, 1998), (Castelfranchi und Falcone, 1999)). Bisher wurden aber weder das mathematische, noch das kognitive Modell empirisch überprüft.

Nach Castelfranchi et al. (1997) ist Vertrauen eine Eigenschaft die einem Charakterzug entspricht:

- Sie unterscheidet einen Agenten von anderen Agenten.
- Ist relativ stabil.
- Ist mental.
- Steht im Zusammenhang mit Motivationen, der Entscheidungsfindung und dem Planen.

Die Autoren betonen die Wichtigkeit der Modellierung von Charakteren, da gerade in einem *Offenen System* nach Hewitt verschiedene Lösungen zu einem Problem, verschiedene Reaktionen zu Situationen, verschiedene Schlußfolgerungsarten und verschiedene Prioritäten der Ziele für eine bessere Problemlösung notwendig sind. Die Modellierung solcher individueller Agentenunterschiede ist ihrer Meinung nach die Modellierung von Charaktereigenschaften und damit auch von Persönlichkeiten.

2.3.3. Zusammenfassung

Es zeigt sich, daß es eine Reihe von verschiedenen Definitionen von Vertrauen gibt, die für eine Vielzahl von Kontexten entworfen wurden. Nur wenige davon eignen sich als Grundlage für eine Formalisierung. Es gibt eine Reihe von Begriffen, die (so hat die bisherige Forschung gezeigt) nur sehr schwer von Vertrauen zu unterscheiden sind. Dazu zählen *Glaubwürdigkeit*, *Verläßlichkeit*, *Vertrauenswürdigkeit* und *Kompetenz*. Folglich muß man in einer Formalisierung sehr genau definieren, auf welche Konzepte man sich bezieht und was ihr Anwendungsgebiet ist. Sonst läuft man Gefahr, mit anderen Arbeiten nicht vergleichbar zu sein und zur Verwirrung beizutragen statt Klarheit zu schaffen. Unabhängig von der Verstricktheit dieser Konzepte kristallisiert sich heraus, daß Vertrauen in direktem Zusammenhang mit den Begriffen *Erwartung, Risiko*, *Macht* und *Kommunikation* stehen muß. Dies zeigt deren konstantes Auftreten in der Literatur über die Wissenschaften hinweg. Eine

Formalisierung von Vertrauen muß einen Bezug zu diesen Begriffen schaffen. Castelfranchi et al. haben eine Theorie beschrieben, die sowohl unter dem Aspekt der Adäquatheit, als auch der Implementierbarkeit sehr erfolgversprechend sind. Der Nutzen von Vertrauen in menschlichen Gesellschaften wird allgemein beschrieben als die Reduktion der Komplexität und die Handhabbarkeit einer (ungewissen) Zukunft durch Orientierung und Stabilisierung. Eine Übertragung auf Multi-Agenten Systeme ist wünschenswert, da auch hier Reduktion der Komplexität und die Handhabbarkeit einer ungewissen Zukunft wichtige Aufgaben für die Konstrukteure sind.

2.4. Entscheidungs- und Spieltheorie

Die Anwendung der Spieltheorie für die Vertrauensforschung ist in der Literatur umstritten. Einerseits hat die Spieltheorie für viele Untersuchungen eine klar strukturierte Basis geboten. Dieses Angebot wurde z.B. von der Psychologie stark genutzt. Andererseits wird kritisiert, daß oft inhärent auf Konfrontation ausgerichtete Spiele untersucht werden. Gegner der Anwendung der Spieltheorie, wie z.B. Marsh, sehen aber zumindest den Vorteil, daß die Spieltheorie Experimentalumgebungen liefert, in denen Interaktionen studiert werden können. Für die Spieltheorie spricht auch, daß es ja gerade Situationen mit möglicher Konfrontation oder in Konflikt stehenden Zielen sind, in denen Vertrauen von zentralem Interesse ist. Dies ermöglicht, wie er zugibt, auch Untersuchungen darüber, inwiefern Vertrauen diese Interaktionen beeinflußt. An dieser Stelle kann nur ein kurzer Überblick in die Bereiche der Entscheidungs- und Spieltheorie gegeben werden. Daher wird im folgenden hauptsächlich auf die in dieser Arbeit verwendeten Konzepte eingegangen.

2.4.1. Grundlagen

Der Grundstein für die Entscheidungs- und Spieltheorie wurde von dem Mathematiker John von Neumann und dem Wirtschaftswissenschaftler Oskar Morgenstern gelegt (von Neumann und Morgenstern, 1944). Diese Theorien sind für Multi-Agenten Systeme sehr interessant, da die Rückführung des Koordinationsproblems auf das Verhandlungsproblem eine elegante Lösung darstellt. Nach Fischer et al. (Fischer et al., 1998) beschäftigt sich die Entscheidungstheorie damit, mathematische Modelle von Entscheidungssituationen zu liefern, diese zu analysieren und Methoden zur Verfügung zu stellen, wie eine Lösung automatisch berechnet werden kann. Die Konzepte der Entscheidungstheorie, die für diesen Prozeß benutzt werden, stammen dabei nicht aus der Psychologie sondern sind aus formalen Eigenschaften des Problems abgeleitet. Außerdem ist die Entscheidungstheorie in der Lage Organisationsstrukturen, Mechanismen zur Ressourcenverteilung und Konfliktauflösung zu analysieren und entwerfen. Sie beschäftigt sich (zumindest im klassischen Sinne) mit der Interaktion eines Agenten mit der meist abstrakt beschriebenen Umwelt. In beiden Theorien geht man von *rationalen* Agenten aus (zu dem Begriff *rational* siehe auch S. 11).

Die Spieltheorie unterscheidet sich dadurch, daß sie die Interaktion mit anderen Agenten und der Abhängigkeit der eigenen Entscheidung von deren Strategien betrachtet. Dies spiegelt sich auch direkt in der Spezifizierung der Situationen wieder,

auf die die Spieltheorie angewendet werden kann. Dazu gehört, daß eine Menge von Spielern (Agenten) an dem Spiel teilnimmt. Außerdem müssen die Spieler gleichzeitig ihre Entscheidung über ihre Handlung (Spielzug) treffen. Währenddessen hat zwar jeder Spieler absolutes Wissen über die Spielregeln, aber nicht über die Spielzüge der anderen Spieler (und damit auch nur eingeschränkt über das eigene Spielergebnis). Wichtig hierbei ist, daß das Spielergebnis des Spielers (auch der *pay-off* genannt) von der Kombination der Spielzüge *aller* Spieler abhängt. Das Spielergebnis wird üblicherweise in Punkten angegeben, die dann verschieden interpretiert werden können. Schließlich kennt jeder Spieler diese Spezifizierung und weiß auch, daß jeder andere Spieler sie kennt. Ein Spiel heißt dabei *in Normalform*, wenn die Menge der Spieler und für jeden Spieler die Menge der Handlungsoptionen (auch *Strategien* genannt) bekannt ist. Die Spieltheorie teilt die Menge der Spiele üblicherweise in 2-Personen und in n-Personen Spiele ein. Eine andere Einteilung ist die Einteilung in Konstantsummenspiele und in Nichtkonstantsummenspiele. Letztere Einteilung bezieht sich auf die Summe der Punkte, die in einem Spiel vergeben werden. Ein Spiel ist ein Konstantsummenspiel, wenn gleichgültig welche Strategien die Spieler wählen, die Summe der Spielergebnisse immer gleich ist. Dies bedeutet, daß schon per Definition des Spiels die Interessen der Spieler im Konflikt stehen. Nicht-konstantsummenspiele bieten die Möglichkeit, daß eine Koordination der Strategien für beide Vorteile bringt.

In der Regel betrachtet die Spieltheorie zunächst *einfache Spiele*, d. h. Spiele die nur aus einem Spielzug bestehen. Wird ein solches Spiel wiederholt hintereinander ausgeführt, dann bezeichnet man dies als iteriertes Spiel. Das wohl berühmteste Beispiel für ein solches iteriertes Spiel ist das von Axelrod durchgeführte Turnier, in dem Agenten mit verschiedenen Strategien gegeneinander das Gefangenendilemma gespielt haben (Axelrod, 1984). Auf das Gefangenendilemma wird im folgenden noch näher eingegangen. In iterierten Spielen wird auch die Evolution von Spielstrategien untersucht. Das Ergebnis einer solchen Untersuchung sind Strategien die evolutionär stabil sind (auf englisch *evolutionary stable strategies* oder *ESS*), d.h. Strategien, die sich gegen neu auftretende Strategien behaupten können und sich nicht selbst auslöschen. Bei der Analyse von optimalem Verhalten in einfachen Spielen hat der Begriff der *Dominanz* große Bedeutung. Die Auswahl einer dominanten Strategie ist für einen Agenten vorteilhaft. Eine Strategie heißt *dominant*, wenn sie für jede Kombination von Strategien, die die anderen Spieler wählen könnten, auf jeden Fall ein gleichhohes, in einem Fall sogar ein höheres Spielergebnis liefert. In diesem Zusammenhang ist auch der Begriff des *Nash-Equilibriums* zu nennen. Eine Kombination von Zügen der teilnehmenden Spieler heißt Nash-Equilibrium, wenn es für keinen der Agenten einen Anreiz gibt von seinem Zug abzuweichen, solange alle anderen ihren Zug beibehalten. Interessanterweise gibt es Spiele wie z.B. das Gefangenendilemma, in denen das Nash-Equilibrium die für alle Spieler zusammengenommen schlechtmöglichste Kombination von Spielzügen ist.

Nach Zeng und Sycara werden in der Entscheidungstheorie die Annahmen gemacht, daß der Agent über vollständiges Wissen über seine eigenen Präferenzen in Form einer *Nutzenfunktion* (*utility function)* und den Wahrscheinlichkeiten, mit denen verschiedene Ereignisse eintreffen, verfügt, was keine trivialen Annahmen sind. Damit wurde das Problem umgangen, Abstufungen einer perfekten Rationalität definieren zu müssen (Zeng und Sycara, 1996). Zeng and Sycara sehen dabei die

Einschränkung, daß sich die Modelle zu sehr auf das Ergebnis einer Verhandlung konzentrieren und nicht so sehr auf den Prozeß. Sie werfen einer Reihe von Ansätzen aus der VKI und der Spieltheorie (im Zentrum der Kritik stehen (Rosenschein und Zlotkin, 1994), (Osborne und Rubinstein, 1994)) vor, daß diese Ansätze nur bedingt realistisch sind, da sie für Koordination und Verhandlung vorausberechnete und spezifische Lösungen angeben und keine Konzepte angeben, wie Agenten sich flexibel an ihr Verhandlungsumfeld anpassen können (Zeng und Sycara, 1996).

Unter den obigen Annahmen wird aber die Berücksichtigung von Strategien, die sich über mehrere Spiele erstrecken, zu einem Problem. Es ist nämlich denkbar, daß der Agent Vorhersagen über die Entscheidungen der anderen machen wird, bei deren Berechnung er allerdings in Betracht ziehen wird, daß diese versuchen, Vorhersagen über sein Verhalten zu machen. Die damit entstehende unendliche rekursive Schachtelung taucht in der Literatur unter dem Namen *outgessing regress* auf (Young, 1975). Um dieses Problem zu umgehen, werden in der Spieltheorie die folgenden drei (restriktiven) Annahmen gemacht:

- Die Anzahl der Spieler und ihre Identität sind fest und allen bekannt.
- Alle Spieler verhalten sich vollständig rational und jeder Spieler weiß dies. Die Menge der Alternativen jedes Spielers ist bekannt und fest.
- Die Risiko-Hemmschwelle jedes Spielers und seine Nutzenfunktion sind fest und allen an der Entscheidung beteiligten Spielern bekannt.

Um die Schwierigkeiten, die mit den strategischen Überlegungen der Agenten bei n-Personen Spielen verknüpft sind in den Griff zu bekommen, wurden detaillierte Entscheidungsregeln eingeführt, die Konzepte wie die relative Macht der Spieler modellieren (zentrale Begriffe sind hier Nash-Equilibrium, Shapley Wert und Pareto-Effizienz). Eine gut verständliche Einführung findet sich in (Fischer et al., 1998).

Um Aufgaben unter Agenten aufzuteilen, liefert die Spieltheorie das Instrument der *Auktion*. Auktionen lehnen sich an die Marktmetapher an. In den Wirtschaftswissenschaften wurde gezeigt, daß Märkte in der Lage sind, Ressourcen mit geringer oder keiner zentraler Rolle effizient zu verteilen, selbst wenn nur unvollständiges Wissen über die Agenten verfügbar ist (Ruß, 1997). An dieser Stelle soll insbesondere auf das *contract net protocol* (*CNP*) eingegangen werden, da es auch im weiteren verwendet wird.

Beim *contract net protocol* handelt es sich um einen Mechanismus zur Verteilung von Gütern oder Dienstleistungen, wie zum Beispiel Aufgaben, einer Ressource oder Kooperationsangeboten. Einer der Agenten (der sogenannte *manager*) verfügt über dieses Gut oder die Dienstleistung. Er befragt eine Menge von Agenten, zu welcher Gegenleistung sie bereit sind. Diese Agenten, auch *Bieter* (*bidders*) genannt, führen ihre eigenen Berechnungen durch, um eine für sie akzeptable Gegenleistung zu berechnen. Das Ergebnis teilen sie dem *manager* mit. Dieser entscheidet sich aufgrund der bei ihm eingegangenen Angebote. Als Erweiterung kann der Angebotsphase ein Zeitlimit, innerhalb dessen die Bieter antworten müssen, zugeordnet sein. Später eingegangene Angebote werden dann nicht mehr berücksichtigt ((Smith, 1980), (Smith, 1982)). Diese Erweiterung ist ein wichtiger Ansatz, um zeitkritische Verteilungen zu realisieren. Dieser Ansatz findet direkten Bezug zu den *beschränkt*

rationalen Agenten. Darüber hinaus bieten spezielle Auktionsprotokolle auch den Austausch von mehreren Gütern oder Dienstleistungen gleichzeitig an. Einen Überblick über Spieltheorie im Kontext der Multi-Agenten Systeme und einer Reihe von Verhandlungen und Aktionen gibt (Fischer et al., 1998). Zur Spieltheorie im allgemeinen gibt es eine ganze Reihe von einführender Literatur. An dieser Stelle seien nur die Bücher von Straffin oder Fudenberg und Tirole erwähnt ((Straffin 1996), (Fudenberg und Tirole 1991)).

2.4.2. Das Gefangenendilemma

Ein Spiel, welches in der Spieltheorie einen besonderen Bekanntheitsgrad erlangt hat, ist das von Luce und Raiffa beschriebene *Gefangenendilemma*[1] (Luce und Raiffa, 1957). Das besondere an diesem Spiel ist, daß sich mit ihm, obwohl es aus sehr einfachen Regeln besteht, sehr komplexe Vorgänge, z.B. gesellschaftlicher Natur, modellieren lassen. Es zeigt, daß individuell rationale und optimale Entscheidungen auf Gesellschaftsebene zu ineffizienten und suboptimalen Lösungen führen können. Außerdem zeigt seine Analyse den Einfluß, den Kommunikation bzw. das Fehlen von Kommunikation auf die möglichen Lösungen hat. Dieser Eigenschaften wegen erfreut es sich auch interdisziplinär einer großen Beliebtheit. Es ist bekannt im Bereich der Multi-Agenten Systeme (Rosenschein, 1985), der Soziologie (Rapoport und Orwant, 1962), der Evolutionsbiologie (Smith, 1982), der Politikwissenschaften (Axelrod, 1984) und der Volkswirtschaftslehre (Daws und Thaler, 1988).

Bei dem Gefangenendilemma handelt es sich um ein Spiel, das zwischen zwei Agenten gespielt wird: Jeder der beiden Spieler (Agent *A* oder *B*) hat zwei Handlungsoptionen wie in Abbildung 3 dargestellt. *C* steht für *cooperation* (Kooperation) und D für *defection* (Verrat). In der Literatur sind zwei Arten von Matrizen bekannt: Matrizen mit Belohnungen und Matrizen die die Höhe von Strafen angeben. Wir betrachten hier ein Gefangenendilemma, bei dem Belohnungen ausgesprochen werden. Kooperieren beide Agenten, erhalten sie jeweils **R** Punkte *(Reward)*, verraten sie sich gegenseitig, erhalten sie jeder nur **U** Punkte *(Uncooperative)*. Spielen sie beide verschieden, erhält der Spieler, der *C* spielt **S** Punkte *(Sucker's Payoff)*, der andere **T** *(Temptation)*[2].

Das Dilemma besteht nun darin, daß zwischen den vier verschiedenen Werten folgende Relationen gelten:

$$T>R>U>S \qquad (6)$$

$$\text{und } 2R>(S+T). \qquad (7)$$

---

[1] Seinen Namen hat das Gefangenendilemma von Albert W. Tucker, einem Mathematikprofessor aus Princeton. Von diesem stammt auch die seither bekannte Geschichte der beiden Gefangenen, die ihr Dilemma „spielen". Konzipiert wurde das Spiel aus mehr theoretischem Interesse von Melvin Dresher und Merrill Flood von der RAND Corporation, die zeigen wollten, daß es Spiele gibt, die ein einziges Nash-Equilibrium besitzen, das aber nicht pareto-optimal ist (Straffin 1996). Eine anekdotenhafte Einführung inklusive der Geschichte von Tucker findet sich in (Watzlawick, 1997).

[2] Die Bezeichnungen der Punktergebnisse sind Rapoport und Chammah (1970) entnommen.

Deshalb können die Agenten zwar durch gegenseitige Kooperation gemeinsam die höchste Punktzahl erreichen (insgesamt *2R*). Der Gewinn jedes einzelnen Agenten ist aber größer, wenn er einen kooperierenden Agenten verrät ($T>R$). Spieltheoretisch gesprochen bedeutet dies, daß *D* für beide Spieler die dominante Strategie ist[1] (für jeden möglichen gegnerischen Zug hat *D* ein höheres Ergebnis als *C*: $T>R$ und $U>S$). Die Konstellation die sich einstellt, wenn beide Spieler ihre dominante Strategie spielen (das Nash-Equilibrium) ist also *D*/*D*. Dies ist jedoch nicht Pareto-optimal (die für beide Spieler günstigste Konstellation). Beide würden die Konstellation *C*/*C* gegenüber *D*/*D* vorziehen.

B / A	C	D
C	**R** **R**	**T** **S**
D	**S** **T**	**U** **U**

Abbildung 3: Abstrakte Ergebnismatrix für das Gefangenendilemma

Es gibt eine Fülle von Literatur in der Informatik, in denen Untersuchungen mit Hilfe des Gefangenendilemmas durchgeführt werden. Dies sind zum Teil theoretische und zum Teil empirische Untersuchungen. Eine Reihe von Publikationen beschäftigt sich damit, wie Agenten entworfen werden können, so daß in großen Gesellschaften in denen das Gefangenendilemma gespielt wird Kooperation als emergentes Verhalten auftritt. Dabei gibt es Ansätze, die Agentenverhalten mit Hilfe von Endlichen Automaten modellieren und diese Automaten durch ihre Agenten aufgrund von Beobachtungen lernen lassen ((Carmel und Markovitch, 1998), (Mor et al., 1996)). Die meisten Arbeiten setzen auf große Mengen von Daten, die dann probabilistisch analysiert werden können ((Biswas et al., 1999b), (Marsh, 1994)) oder gewähren den Agenten Wissen über die Strategien anderer ((Armstrong und Durfee, 1998), (Bazzan et al., 1997)).

Abschließend sei angemerkt, daß die Wahl der Strategie *C* trotz der Tatsache, daß das Maximum unter der Wahl von *C* kleiner als das Maximum unter der Wahl von *D* ist, genau das ist, was Deutsch als eine *vertrauensvolle Entscheidung* bezeichnet (Deutsch, 1973).

---

[1] Die Werte selbst spielen für die Analyse eine untergeordnete Rolle. Solange sie diese Ungleichungen erfüllen bleibt die zentrale Eigenschaft des Gefangenendilemmas, die Ungleichheit von Pareto-Optimum und Nash-Equilibrium erhalten. In der Literatur verwendete Werte sind z. B. (5, 3, 1, 0) für T, R, U, S. Aus dem „Axiom der linearen Invarianz" von Nash (1950) bezüglich der Lösung von Nichtnullsummenspielen folgt, daß jede Transformation durch eine positive lineare Funktion lösungserhaltend ist.

Kapitel 3
Problemstellung

„Doveryay, no proveryay."
(Vertraue, aber überprüfe.)

Ronald Reagan, mit einem Lenin-Zitat bei der Unterzeichnung
des Mittelstreckenwaffen-Abkommens 1987

Zunächst wird die in dieser Arbeit behandelte Problemstellung beschrieben. Mit Verweis auf das vorhergehende Kapitel wird gezeigt, inwiefern dieses Problem noch nicht in der Literatur behandelt wurde. Der zweite Abschnitt geht auf spezielle Schwierigkeiten ein, die bei der Lösung des Problems zu beachten sind. Schließlich betont der dritte Abschnitt die praktische Relevanz dieses Problems anhand der Beschreibung von vier aktuell in der Forschung behandelten Anwendungsszenarien von Vertrauen in Multi-Agenten Systeme.

3.1. Behandelte Problemstellung

Agenten sehen sich in Interaktionen mit anderen Agenten in *Offenen Systemen* der Gefahr ausgesetzt, daß ihre Interaktionspartner nicht benevolent sind. In vielen Anwendungen sind sie aber auf die Interaktion mit anderen angewiesen (siehe Abschnitt 3.3). Daher ist es für sie wichtig herauszufinden, welche Agenten für sie zur Kooperation geeignet sind bzw. welche Agenten nur darauf warten, einen anderen auszubeuten. Dies zu bestimmen ist im allgemeinen sehr schwierig. Diese Aufgabe wird zusätzlich erschwert, wenn in einem solchen System nur wenige (eventuell sehr wichtige) Interaktionen stattfinden, in denen Erfahrungen gesammelt werden können, oder ein in diesem System neuer Agent noch nicht die Gelegenheit hatte, viele Interaktionen zu beobachten.

Diese Arbeit soll drei Ziele erreichen. Erstens soll eine *Theorie* entwickelt werden, wie Agenten, trotz weniger Daten über das Verhalten anderer Agenten, qualifizierte Entscheidungen über die Wahl ihrer Interaktionspartner treffen können. Um dies zu erreichen, sollen sie die Kommunikation mit anderen (möglicherweise betrügerischen) Agenten nutzen. Wie im vorherigen Kapitel dargelegt, konzentrieren sich bisherige Ansätze darauf, nicht das Verhalten von Agenten einzuschätzen,

sondern sie versuchen die genaue Strategie anderer Agenten in Form von Endlichen Automaten zu bestimmen oder aber benutzen probabilistische Ansätze, die große Mengen an Beobachtungen benötigen. Die Voraussetzung der Verfügbarkeit vieler Daten über das Verhalten von fremden Agenten lehnen wir für die im folgenden beschriebenen Anwendungen als unrealistisch ab. Ebenso können wir nicht davon ausgehen, daß andere Agenten Einblicke in ihre Strategien erlauben. Die im vorigen Kapitel beschriebenen Arbeiten gehen folglich von Voraussetzungen aus, die nur in benevolenten Systemen oder theoretischen Spielen gegeben sind. In der Praxis kann ein Agent jedoch nicht viele (eventuell negative) Erfahrungen riskieren, nur um seine Modelle von anderen zu erstellen. Um dies zu erreichen, wird ein Konzept von *Vertrauen* erarbeitet, daß es den Agenten erlaubt, Schlußfolgerungen über andere Agenten zu ziehen. Es ermöglicht ihnen mit anderen Agenten zu kommunizieren und sie über das Verhalten anderer zu befragen und die Ehrlichkeit dieser Aussagen zu bewerten. Insbesondere sind sie dadurch in der Lage auch mit betrügerischen Agenten, die ihnen verfälschte Daten schicken, umzugehen. Damit können sie dann ihre Wissensbasis erweitern (falls sie die Aussagen für ehrlich halten). In gewisser Weise sollen diese Agenten also „sozial kompetenter" sein. Die bisher in der Literatur dargestellten Konzepte von Vertrauen, wie etwa von Deutsch und Luhmann, eignen sich nicht zur Formalisierung, da sie zu unscharf oder widersprüchlich sind (siehe Abschnitt 2.3.1). Da eine Formalisierung eine wichtige Voraussetzung für den Einsatz als Agententechnologie ist, müssen die Konzepte zunächst formaler, unter Umständen mit weniger Facetten in ihrer Bedeutung, beschrieben werden.

Andere Konzepte, wie in Abschnitt 2.3.2 beschrieben, benötigen weitere Verfeinerung. Das Konzept von Marsh setzt Vertrauen implizit mit Kooperationswillen gleich (wir zeigen in Abschnitt 4.1.4, daß dies eine unzureichende Modellierung ist). Außerdem können die Agenten, mit denen er seine Theorie getestet hat, nicht beide ihren Interaktionspartner aussuchen. Ein Agent zwingt einem anderen eine Interaktion auf, wodurch bei der empirischen Analyse ein Teil dem Zufall überlassen bleibt. Schließlich gibt es in seiner Analyse keinerlei Kommunikation zwischen den Agenten, in der sie sich über ihre Intentionen austauschen könnten. Obwohl er Vertrauen modelliert, bleibt so die Möglichkeit des Betruges unberücksichtigt. Dadurch ist sein Modell nicht allgemein genug, als daß es für eine Bewertung von „Zeugenaussagen" geeignet wäre. Zwar hat er in seiner Doktorarbeit versucht, eine viel größere Theorie zu entwerfen. Er hat aber sehr viele Facetten des Problems nur anreißen können. Daher soll in dieser Arbeit nur eine der drei Ebenen in seiner Theorie untersucht werden, nämlich das Vertrauen in andere Agenten. Die Arbeiten von Castelfranchi et al. haben einen anderen Schwerpunkt. Sie gehen davon aus, daß die Wahrscheinlichkeit für die Ehrlichkeit der Intention eines anderen Agenten eingeschätzt werden kann, ohne zu erklären, wie dies berechnet werden kann. Außerdem gehen sie nicht darauf ein, was es bedeutet, wenn während der Kommunikation betrogen wird. Sie berechnen zwar die Wahrscheinlichkeit dafür, daß eine Handlung eines Agenten (also auch Kommunikation) nicht vertrauenswürdig ist. Daraus kann ein Agent aber noch nichts über den Inhalt der Kommunikation folgern. Dieser Mangel soll im folgenden behoben werden.

In dieser Arbeit werden auf der theoretischen Seite zum Erreichen des ersten Ziels zwei Dinge behandelt: 1. wie kann die Wahrscheinlichkeit der Einhaltung der Intention durch Beobachtung bestimmt werden und 2. wie kann die Kommunikation

mit (möglicherweise) betrügerischen Agenten dazu dienen, dieses Modell schneller zu erhalten und exakter zu machen. Wie schon im vorigen Kapitel gezeigt, wurde die Kommunikation mit anderen Agenten bisher in der Literatur nur für den Fall der benevolenten Agenten behandelt (siehe z.B. (Biswas et al., 1999a), (Zacharia, 1999), (Xiang, 1994)).

Um die hier vorgestellte Theorie zu überprüfen, wird eine *Experimentalumgebung* entworfen, die auf Arbeiten der Spieltheorie aufbaut und eine Popper'sche Falsifizierung erlaubt. Diese Experimentalumgebung ist das zweite Ziel dieser Arbeit. Schließlich wird eine *Implementierung* der Theorie, des Testbetts und einfacher Agenten vorgestellt. Das dritte Ziel ist die *Evaluation* dieser Implementierung. In dieser Evaluation soll bestimmt werden, wie die Performanz der Agenten von Parametern der Szenarien abhängt und welche Performanzänderung unter Verwendung der vorgestellten Theorie in Abhängigkeit von diesen Parametern zu erwarten ist.

3.2. Spezielle Herausforderungen

Bei der Modellierung von Vertrauen sind zwei Schlüsselfragen zu beantworten. Zum einen, wie verändert sich das Vertrauen in einen Agenten, wenn es nicht selbst beobachtet wurde, sondern Agentenverhalten nur durch „Hörensagen" erfahren wurde. Zum anderen, wie können mehrere solcher Aussagen miteinander kombiniert werden. Beide Fragen werden kontrovers in der Literatur besprochen und sind in den folgenden beiden Abschnitten diskutiert.

3.2.1. Warum ist Vertrauen nicht transitiv?

Zunächst sollte festgehalten werden, daß es sich bei Vertrauen, oder genauer der Vertrauenswürdigkeit, nicht um eine Eigenschaft handelt, die in einem System jedem Beobachter bekannt ist. Es handelt sich dabei vielmehr um individuelle, möglicherweise falsche Annahmen über das Verhalten anderer Akteure. Solche Annahmen werden auch als *belief* bezeichnet. Es ist gerade das Interessante an dem Phänomen Vertrauen, daß ein Akteur einem anderen vertraut, obwohl dieser möglicherweise gar nicht vertrauenswürdig ist.

Aus der menschlichen Intuition ergibt sich die Vermutung, daß das Konzept der Vertrauenswürdigkeit eine Eigenschaft besitzt, die allgemein als *Transitivität* bezeichnet wird. Überträgt man diese Eigenschaft auf den *belief* der Vertrauenswürdigkeit, ergibt sich eine Schlußfolgerung, die in Anlehnung an die Syntax der Prädikatenlogik folgendermaßen beschrieben werden könnte:

$$belief_A(B\ vertrauenswürdig) \wedge belief_B(C\ vertrauenswürdig) \supset belief_A(C\ vertrauenswürdig). \quad (8)$$

Ein *belief* ist jedoch nicht immer nur eine Aussage, die sich in zweiwertiger Logik ausdrücken läßt. Ein wesentlich adäquateres Modell ist es, Vertrauen als eine Wahrscheinlichkeit zu modellieren, mit der ein bestimmtes Ereignis eintritt. Diese Vorgehensweise entspricht nicht nur besser der Intuition, sie ist auch in der Literatur unbestritten. Betrachen wir also einen Agenten *B* aus der Perspektive eines Agenten *A*. Für den Fall, daß *B* absolut vertrauenswürdig ist, hat obige Formel sicher etwas für sich.

Angenommen, *B* ist *nicht* hundertprozentig vertrauenswürdig und *A* weiß dies. Was bedeutet dann die Tatsache, daß *B* z.B. siebzigprozentiges Vertrauen in *C* hat? Um die Sache noch komplizierter zu machen: Wenn B nicht hundertprozentig vertrauenswürdig ist, warum sollten wir davon ausgehen, daß seine Aussage, daß er *B* mit einer Wahrscheinlichkeit von siebzig Prozent vertraut, ehrlich ist? Ist es nicht naheliegender, daß er bei dieser Aussage gelogen hat? Und wenn er gelogen hat, in welcher Hinsicht hat er dann gelogen?

Ein Ansatz der sich häufig in der Literatur findet, besteht darin, diese beiden Wahrscheinlichkeiten miteinander zu „gewichten", also den *belief* als Produkt aus der Wahrscheinlichkeit der Vertrauenswürdigkeit des Zeugen *B* und dessen Aussage zu nehmen. Diese Idee ist genauso intuitiv wie falsch, wie z. B. Maurer darlegt. Wäre das nämlich eine korrekte Berechnung, so würde dies implizieren (und das folgt zwingend aus der Wahrscheinlichkeitstheorie), daß beide Wahrscheinlichkeiten voneinander *unabhängig* sind. Gerade das ist aber im allgemeinen falsch. Angenommen, die Aussage des Zeugen wäre unabhängig von seiner eigenen Vertrauenswürdigkeit, so bräuchte man die Vertrauenswürdigkeit des Zeugen ja gar nicht zu betrachten (Maurer, 1996). Eine Darstellung, warum Vertrauen selbst als zweiwertiges Konzept nicht transitiv sein kann, findet sich in (Christianson und Harbison, 1997).

Ist die Modellierung von Vertrauen in Zeugen das Ziel, muß also in Betracht gezogen werden, daß die Zeugen auch *nicht vertrauenswürdig* sein können und daß ihre Aussagen dementsprechend nicht (unbedingt) der Wahrheit entsprechen. Ein Ansatz der dies leisten will, sollte berücksichtigen, daß Vertrauen mehr als nur zwei triviale Zustände annehmen kann, und eine detailliertere Berechnung dieser Eigenschaft zulassen. Weiterhin muß er berücksichtigen, daß im günstigsten Fall mehrere Zeugenaussagen über einen Agenten zur Verfügung stehen und daß diese korrekt miteinander verrechnet werden müssen. Dies stellt ein weiteres nicht-triviales Problem dar, welches im nächsten Abschnitt dargelegt wird.

3.2.2. Kombinieren mehrerer Zeugenaussagen

Stehen einem Agenten Aussagen von mehreren Zeugen zur Verfügung, so stellt sich die Frage, wie diese Aussagen miteinander kombiniert werden. Lassen wir zunächst das im vorigen Abschnitt diskutierte Problem der Vertrauens(un)würdigkeit der Zeugen außer acht. Nehmen wir an, es liegen zwei Zeugenaussagen vor (die Vorgehensweise muß auf *n* Zeugen abstrahiert werden können). Nehmen wir weiter an, beide Zeugen berichten ein zu sechzig Prozent vertrauenswürdiges Verhalten von Agent *C*. Wie soll nun der Agent, der diese beiden Aussagen hat, *C* einschätzen? Die offensichtliche Antwort, nämlich das Mittel der beiden Aussagen, also wiederum sechzig Prozent, ist nicht korrekt. Bei dem Mitteln der beiden Aussagen wird nämlich deren mögliche gegenseitige Abhängigkeit außer acht gelassen.

Der Effekt, den diese gegenseitige Abhängigkeit haben kann, ist in Abbildung 4 dargestellt. Beide Zeugen *Z1* und *Z2* haben jeweils einen Ausschnitt des Verhaltens von *C* beobachtet. Ein „√" bedeutet in der Abbildung die (mitgeteilte) Beobachtung eines ehrlichen Verhaltens, ein „X" ist eine Beobachtung eines unehrlichen Verhaltens. Im günstigsten Fall, wenn sie also beide absolut ehrlich über ihre Beobachtungen sind, würde das Mitteln beider Aussagen eine Einschätzung ergeben,

die nicht der Einschätzung entspricht, die möglich gewesen wäre, wenn beide auch gesagt hätten, welchen Abschnitt des Verhaltens sie beobachtet haben. Es ergibt sich sogar zwischen dem Verhalten (zu 42,86% ehrlich) und dem Mittel der Aussagen (sechzig Prozent) ein relativer Fehler von fast vierzig Prozent.

Spieldaten															Durchschnitt
Verhalten von *C*	√	√	√	√	√	√	x	x	x	x	x	x	x	x	$\frac{6}{14}$ **= 42,86%**
Aussage Z1	√	√	√	√	√	√					x	x	x	x	$\frac{6}{10}$ **= 60%**
Aussage Z2	√	√	√	√	√	√	x	x	x	x					$\frac{6}{10}$ **= 60%**

Verhalten: **42,86%** Zeugenaussagen: **60%** Relativer Fehler: **39,99%**

Abbildung 4: Inadäquates Kombinieren von Zeugenaussagen

Der Hinweis auf dieses Problem ist nicht trivial, da es z. B. in (Zacharia, 1999), (Beth et al., 1994) und in den Akkumulatoren von (Armstrong und Durfee, 1998) nicht beachtet wird. Da die Vernachlässigung dieses Aspekts möglicherweise größere Fehler in der Einschätzung bewirkt, als selbst der ausgefeilteste Lernalgorithmus je wieder wettmachen kann, ist es sehr wichtig, daß ein Zeuge auch mitteilt, welchen Teil des Verhaltens er beobachtet und daß diese Information bei der Berechnung beachtet wird.

3.3. Praktische Anwendungen

Mittlerweile gibt es auch in der VKI eine reichhaltige Literatur zum Thema Vertrauen, die das akademische Interesse an diesem Thema belegt. In diesem Abschnitt gehen wir darauf ein, inwiefern der in dieser Arbeit dargestellte Ansatz auch von praktischer Relevanz ist und in welcher Hinsicht es sinnvoll ist, dieser Arbeit nachzugehen

3.3.1. Virtuelle Märkte - Electronic Commerce

Nach Sandholm (1998) werden Verhandlungssyteme mit Agenten, die ihre eigenen Interessen vertreten, immer wichtiger. Dies liegt zum einen an der wachsenden Standardisierung der Infrastruktur für Kommunikation, zum anderen an der wachsenden Attraktivität von virtuellen Unternehmen. Multi-Agenten Technologie unterstützt Verhandlungen auf einer operativen Entscheidungsebene. Die Automatisierung spart dabei Arbeitszeit menschlicher Händler und kann aufgrund der Exaktheit ihrer Berechnungen in strategisch und kombinatorisch komplexen Szenarien

überlegen sein. Wie könnte das Problem für die Bestimmung von günstigen Kooperationspartnern aussehen?

Ganz konkret bietet sich ein Szenario an, wie es im Projekt TeleTruck CC am DFKI bearbeitet wird (Bürckert et al., 2000). In diesem Projekt wird die Zusammenarbeit von Speditionen betrachtet, die in starker gegenseitiger Konkurrenz stehen. Trotzdem wäre es für sie von Vorteil, wenn sie kostenaufwendige Aufträge mit anderen Firmen tauschen würden, die diese Aufträge durch Zusammenlegen mit schon vorhandenen Aufträgen günstiger erledigen können. Trotzdem bearbeiten Speditionen selbst ungünstige Aufträge, da sie sich nicht sicher sein können, daß die Abgabe eines Auftrages an ihren Konkurrenten ihnen in Zukunft im Austausch einen neuen Auftrag bringt. Die Vertrauenswürdigkeit eines Kooperationspartners ist also nicht sicher. Dieses Problem existiert in vielen Bereichen der gewerblichen Zusammenarbeit ganz analog. Daher wollen wir das Problem der Speditionen und dem Auftragsaustausch auf *Agenten* und *Kooperation* verallgemeinern.

Betrachten wir einen Agenten *X*, der mit einem anderen Agenten eine Kooperation eingehen möchte. Diese Kooperation kann im Kauf einer Ware, einer Dienstleistung oder in der komplexeren Vereinbarung einer Zusammenarbeit bestehen. Angenommen der Agent veröffentlicht sein Interesse an einer spezifizierten Kooperation und es melden sich eine Reihe von Agenten, die bereit sind, diese Kooperation mit ihm einzugehen (in unserem Speditionsszenario wären dies Konkurrenten, die einen Auftragstausch anbieten). Von diesen Agenten kennt er einige aus eigener Erfahrung, andere wiederum kennt er nicht. Es ist nun denkbar, daß er über eine Vielzahl von Informationskanälen Zugang zu Informationen über andere Agenten hat. Diese Informationskanäle können sehr verschieden aussehen. Agenten in der Geschäftswelt beziehen ihr Wissen über andere durch (bekannte, befreundete) Geschäftspartner, Zeitungen, Agenturen und allgemein durch die Presse. Betrachtet man das Geschehen an der Börse, ist es leicht einzusehen, daß von diesen Informationen sehr viel abhängt und die Informationen auch nicht notwendigerweise nur aus uneigennützigen Motiven verbreitet werden.

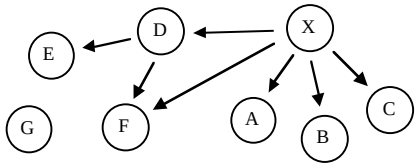

Abbildung 5: Modell der Informationen über Akteure im *Electronic Commerce*

Ein Netzwerk solcher Informationskanäle ist abstrakt in Abbildung 5 dargestellt. Dabei stehen die Knoten für Agenten, die Kanten drücken aus, daß ein Agent etwas über den Knoten, zu dem er eine ausgehende Kante hat, weiß. Agent *X* weiß z.B. etwas über die Agenten *A* bis *D* und *F*, jedoch nichts über *E* und *G*. Angenommen, *X* erhält von allen Agenten wie oben beschrieben ein Angebot. Dann wird er zunächst aufgrund seiner Erfahrungen beurteilen, für wie geeignet er jeden von ihnen als Kooperationspartner hält. Wie aber kann er eine Evaluation von *E* und *G* erhalten? Der Graph in der Abbildung deutet schon an, daß es aussichtslos erscheint, Informationen über *G* zu erhalten. Er zeigt aber auch auf, daß es möglich wäre, Informationen über *E* durch Kommunikation mit *D* zu erhalten. Je wichtiger die

Kooperation mit *E* für *X* ist, desto wichtiger ist es, daß er diese Möglichkeit der Informationsgewinnung nicht außer acht läßt.

Es gibt noch einen weiteren Nutzen dieses Mechanismus, nämlich die Verifikation des eigenen Wissens: In der gezeigten Abbildung hat *X* zwar schon Informationen über *F*. Es ist aber durchaus denkbar, daß diese Informationen noch nicht genug sind, um *F* wirklich gut einschätzen zu können. Auch dann wäre Kommunikation mit *D* sehr wichtig.

3.3.2. Public Key Management

Ein Problem im Bereich der Kryptographie, dem sich in letzter Zeit immer mehr Arbeiten widmen, ist die Schlüsselverwaltung in asymmetrischen Schlüsselsystemen (*public key management*). Das Problem besteht darin festzustellen, ob der Absender einer Nachricht auch der ist, der er vorgibt, zu sein. Um dies zu beweisen, unterschreibt der Absender seine Nachricht digital. Der Empfänger kann unter Verwendung des öffentlichen Schlüssels (*public key*) des Absenders überprüfen, ob die Unterschrift echt ist. Eine der größten Sicherheitslücken bei diesem System, so der *PGP*-Author Phillip Zimmermann, ist die Gefahr, daß sich ein „Spion" in die Datenübermittlung einklinkt und die öffentlichen Schlüssel so manipuliert, daß sie eine für ihn offene Hintertür enthalten. Das Problem des Empfängers besteht nun darin, herauszufinden, ob der öffentliche Schlüssel, den er vom Absender besitzt, auch wirklich der echte ist. An dieser Stelle kommt nun Vertrauen ins Spiel. Es ist vorstellbar, daß in einem Netzwerk von Benutzern der Kryptographie Akteure existieren, denen der Empfänger vertraut und die den öffentlichen Schlüssel des Absenders kennen. Hat der Empfänger die Möglichkeit, diese zu befragen, so kann er eine Überprüfung er Echtheit sofort durchführen. Die Vertrauenswürdigkeit der anderen Akteure kann er wie im vorherigen Beispiel durch Bekanntschaft oder aber durch die Kommunikation mit anderen Akteuren einschätzen. Ein Verfahren, daß unter Menschen für die Ermittlung von Vertrauenswürdigkeit gebräuchlich ist.

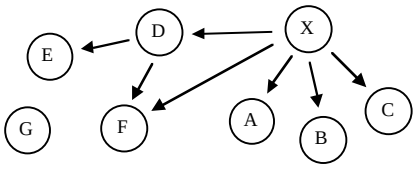

Abbildung 6: Modell der Information über Akteure beim *public key management*

In der Literatur gibt es bereits das Konzept der *Trusted Third Parties*, die die Rolle solcher vertrauenswürdiger „Informanten" übernehmen sollen. An den *Trusted Third Parties* gibt es zum einen ein wirtschaftliches Interesse: das Monopolisieren dieser Daten, um damit Geld zu verdienen. Zum anderen gibt es ein politisches Interesse, diese Daten zentral verfügbar zu haben, beispielsweise für Nachrichtendienste. Beides sind Intentionen, die den meisten Benutzern der Kryptographie eher suspekt sind. Es gibt schon Arbeiten, die einen technischen Rahmen dafür anbieten, um die Vertrauensberechnung dezentral (also durch den Erfahrungsaustausch mit anderen Benutzern) durchzuführen. Deren Vertrauensberechnung berücksichtigt allerdings nicht die in Abschnitt 3.2 vorgestellten Probleme und weist durch die damit einher-

gehende Ungenauigkeit der Berechnung erhebliche Sicherheitslücken auf ((Beth et al., 1994), (Maurer, 1996)).

Dieses Problem läßt sich in direkter Analogie zum vorherigen Szenario beschreiben und wir können daher den gleichen Vertrauensgraphen wie im vorherigen Beispiel benutzen. Lediglich die Kanten haben hier eine andere Semantik: Sie drücken Wissen über Vertrauenswürdigkeit und öffentliche Schlüssel aus (der einfacheren Darstellung wegen gehen wir davon aus, daß niemals nur eine Art von Wissen zur Verfügung steht). Betrachtet man Agent *X* in Abbildung 6, so läßt sich leicht erkennen, daß er keine Aussage über den öffentlichen Schlüssel von *E* machen kann. Angenommen *D* kennt diesen Schlüssel. Dann wäre es ungeschickt, wenn *X* sein Wissen über die Vertrauenswürdigkeit von *D* nicht nutzen könnte, um diese Wissenslücke zu schließen. Außerdem sollte er das Wissen von *D* über *F* nicht außer acht lassen. Möglicherweise hat *X* einen falschen Schlüssel von *F* und nur ein Wissensaustausch mit *D* könnte diese Sicherheitslücke aufdecken.

3.3.3. Mobile Agenten

Ein Problem, das sich direkt mit den vorhergehenden vergleichen läßt, ist das der „mobilen Agenten". Dabei handelt es sich um die Risiken für einen Agenten, der sich von Rechner zu Rechner bewegt und dort seine Berechnungen ausführen will. Diese Rechner könnten den Programmcode oder die Daten des Agenten ausspähen oder manipulieren. Ebenso könnten diese Rechner seine Kommunikation mit anderen Agenten ausspähen oder manipulieren wollen. Ein weiteres Risiko besteht darin, daß sie versuchen könnten, den Programmcode des Agenten falsch auszuführen. Dieses Problem ist in der wachsenden Informationsgesellschaft von höchstem Interesse, so daß sich eine Reihe von Arbeiten damit beschäftigen. Eine Übersicht über den aktuellen Stand der Forschung gibt der von Giovanni Vigna herausgegebene Band (Vigna, 1998).

Bisherige Lösungsansätze konzentrieren sich auf die Verschlüsselung des Codes *(proof-carrying code)* und *black-box* Mechanismen, die dem Agenten eine sichere Umgebung bieten bzw. die Authentifizierung dieser Rechner. Diese „sicheren Hüllen", die um Agenten gelegt werden, sind zum Teil aber sehr rechenaufwendig. Es stellt sich die Frage, ob es nicht Bereiche gibt, in denen die völlige Sicherheit zugunsten schnellerer Antwortzeiten vernachlässigt werden kann (oder muß). An dieser Stelle könnte ein Vertrauensmechanismus helfen, sehr schnell möglichst zuverlässige Rechner zu finden.

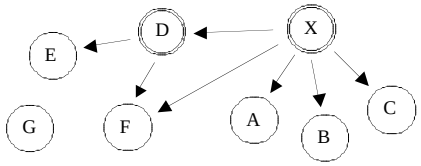

Abbildung 7: Modell der Information über Rechner beim Mobile-Agenten Problem

Wie in den beiden vorhergehenden Szenarien betrachten wir einen Agenten *X* (Abbildung 7). In diesem Szenario stellt *D* einen Agenten dar, der über seine Erfahrungen mit anderen Rechnern berichten kann (beide Agenten sind durch einen doppelten Kreis in der Darstellung gekennzeichnet). Die anderen Knoten repräsen-

tieren Rechner, auf die *X* migrieren könnte. Dabei muß berücksichtigt werden, daß Rechner in der Lage sind Agenten so zu manipulieren, daß sie falsche Aussagen über andere Rechner machen. Es muß also von jedem Agenten die Vertrauenswürdigkeit anderer berücksichtigt werden.

3.3.4. Message-Routing im Internet

Eine gut dokumentierte Sicherheitslücke im Internet ist die Tatsache, daß ein Absender einer Nachricht nicht weiß, über welche Route seine Datenpakete verschickt werden. Dieses Problem wird zur Zeit durch Verschlüsselung der Daten gelöst (siehe dazu auch Abschnitt 3.3.2). Es wäre aber auch denkbar, daß ein Absender sich eine Route aus den Rechnern, von denen er glaubt, daß sie vertrauenswürdig sind, selbst zusammenstellen darf. Jeder Rechner in dieser Route würde dann aus der Nachricht ersehen können, an wen er diese weiterschicken soll.

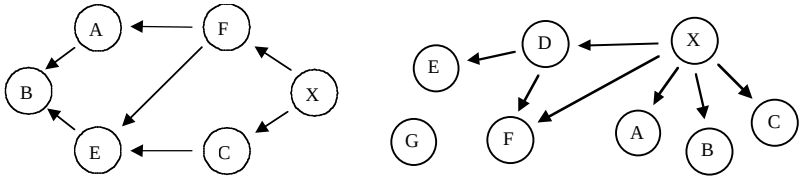

Abbildung 8: Anwendung Message-Routing im Internet

Der rechte Teil in Abbildung 8 ist gestaltet wie in den vorhergehenden Szenarien. Sie zeigt einen Graph mit Kanten, die für Aussagen über Vertrauenswürdigkeit zwischen Agenten stehen. Die Agenten sind gleichzeitig Rechner im Internet, die für die Verschickung von Datenpaketen zuständig sind. Im linken Teil der Abbildung ist eine mögliche Vernetzungstopologie dieser Rechner angegeben. Der Agent *X*, Absender einer Nachricht, kennt diese Topologie und muß nun sein Modell über die Vertrauenswürdigkeit der anderen Agenten benutzen, um eine sichere Route zu seinem Ziel, dem Knoten *B*, zu planen. Es bieten sich drei Routen an. Die Sicherheit der in der Abbildung unteren Route läßt sich ohne Evaluation der Vertrauenswürdigkeit von *D* und dessen Aussage über *E* nicht berechnen. Die Sicherheit der oberen Route könnte er zwar berechnen, die Auskunft von *D* über *F* würde ihm aber weitere Daten geben, auf die er nicht verzichten sollte. Erweist sich der Rechner *A* als nicht vertrauenswürdig und erwägt *X* die Route mit der Verbindung von *F* nach *E*, so ist er doppelt auf die Aussagen von *D* angewiesen. Ohne eine Evaluation der Vertrauenswürdigkeit von *D* ist die Sicherheit dieser Route kaum zu bestimmen.

3.3.5. Zusammenfassung

Wie sich schon an der Ähnlichkeit der Abbildungen gezeigt hat, haben die Szenarien große Gemeinsamkeiten. Vertrauen ist immer dann ein wirksamer Mechanismus für einen Agenten, wenn:

- er anderen Agenten begegnet, deren Intentionen die Verletzung seiner Ziele beinhalten kann.
- er nur sehr wenig Wissen über das Verhalten dieser Agenten hat.

- der Austausch mit anderen dieses Wissen vergrößern kann, eine qualifiziertere Entscheidung ermöglicht und damit den möglichen Schaden minimiert.
- ein exploratives Verhalten, das neue Daten über das Verhalten anderer erzeugen würde, zu teuer wäre oder aus anderen Gründen (Sicherheit) nicht in Frage kommt.
- der Agent durch den Bedarf dieser Daten in Abhängigkeit von anderen gerät.

Im folgenden Kapitel soll ein Testbett für solche Situationen vorgestellt werden, in dem Strategien eingesetzt und evaluiert werden können. Außerdem wird eine Formalisierung von Vertrauen vorgestellt, die eine Lösung für die beschriebenen Szenarien anbietet.

Kapitel 4
Formalismus und Experimentalumgebung

„The essential feature of the [prisoner´s dilemma] is that there is no possibility for "rational" individual behaviour in it, unless the conditions for mutual *trust* exist."

Morton Deutsch, 1973.

Soziale Intelligenz und Interaktion wird in letzer Zeit nicht nur von Soziologen analysiert und beschrieben. Soziale Interaktions- und Organisationssysteme werden zunehmend im Bereich der VKI, insbesondere zur Modellierung von Multi-Agenten Systemen, verwendet. Dabei geht es beim Modellieren von sozialen Prozessen innerhalb von künstlichen Gesellschaften, eben Multi-Agenten Systemen, nicht darum, soziologische Fachbegriffe neu zu definieren, sondern diese als Beschreibungsmetaphern für künstliche Gesellschaften zu nutzen. Solche Metaphern werden beispielsweise von Bazzan, Bordini und Campbell (1997) eingesetzt um zu untersuchen, ob und wie rationales Verhalten in einer künstlichen Gesellschaft von Vorteil ist. Dazu nutzen sie das spieltheoretisch gut untersuchte *Gefangenendilemma (prisoner's dilemma)*. Dieses stellt eine Formalisierung des Konfliktes zweier Individuen zwischen gegenseitiger Unterstützung (Kooperation) oder egoistischer Nutzung von Vorteilen (Verrat) zur Verfügung. Beim sogenannten *iterierten Gefangenendilemma (iterated prisoner's dilemma - IPD)* hat sich gezeigt, daß die rationale egoistische Vorteilsnutzung auf lange Sicht nicht gewinnbringend wirkt (Axelrod, 1984).

Bazzan et al. (1997) haben das *IPD* um soziales Verhalten erweitert und total-rationale Egoisten gegen großzügige Altruisten spielen lassen. Dabei war den Spielern die soziale Rolle des Gegenspielers bekannt. Ihre Ergebnisse deuten darauf hin, daß pure Rationalität im Sinne von immer „Verrat" zu spielen sich für den Egoisten und dessen soziale Gruppe auf lange Sicht nicht positiv auswirkt. Die vorliegende Arbeit erweitert diesen Ansatz in dreierlei Hinsicht. Erstens wird den Agenten das Wissen über das zukünftige Verhalten ihrer Spielpartner nicht mitgegeben, sondern sie lernen dieses anhand ihrer Beobachtungen. Zweitens verhalten sich Agenten nicht entweder egoistisch oder altruistisch, sondern es treten beide Verhaltensweisen

jeweils mit einer gewissen Wahrscheinlichkeit auf. Damit wird die strikte Trennung dieser Verhaltensweisen durch weiche Persönlichkeitsprofile ersetzt. Egoisten können sich in manchen Spielen wie Altruisten verhalten und umgekehrt. Dadurch soll ein realistischeres Spielverhalten modelliert werden. Drittens können die Agenten andere Agenten über deren Beobachtungen befragen. Dies ist insbesondere deshalb wichtig, da die für exakte Einschätzungen essentiell wichtigen Beobachtungen nur sehr eingeschränkt verfügbar gemacht werden. Dabei können die Agenten sogar mit Zeugen umgehen, die bezüglich ihrer wahren Beobachtungen die Unwahrheit sagen. Diese Kompetenz erhalten sie durch ein Vertrauensmodell, das in diesem Kapitel vorgestellt wird. Um die Vertrauenswürdigkeit in potentielle Interaktionspartner einschätzen zu können wird das kognitive Vertrauensmodell von Castelfranchi und Falcone (1998) in einer für unsere Zwecke präzisierten Form genutzt. Dieses Modell zeichnet sich dadurch aus, daß es zugleich qualitative und quantitative Aussagen über Vertrauen macht.

Im ersten Abschnitt präzisieren wir die für die Anwendungsszenarien wichtigen Konzepte Ehrlichkeit, Altruismus und Vertrauen anhand formaler Definitionen. Wir leiten an dieser Stelle spezielle Gleichungen her, nach denen wir im nächsten Kapitel Vertrauen berechnen werden. Im zweiten Abschnitt entwerfen wir ein spieltheoretisches Modell, das *Offen Gespielte Gefangenendilemma mit Partnerwahl.* Dieses Modell abstrahiert von den Szenarien und dient der weiteren Untersuchung als Experimentalumgebung, in der Agenten und Strategien für die Anwendungsszenarien evaluiert werden können. Diese Experimentalumgebung wurde im ersten Workshop „Sozionik" auf der KI-Jahrestagung 1998 vorgestellt (Schillo und Funk, 1998). Sie ist die notwendige Voraussetzung für die Analyse im Kapitel 6. Der letzte Abschnitt geht abschließend auf die verwendeten soziologischen Begriffe ein.

4.1. Formalisierung von Vertrauen

Im folgenden werden die beiden benutzten Modelle von Vertrauen detailliert beschrieben. Zum einen wird das Vertrauen in Kooperationspartner, wie von Castelfranchi und Falcone vorgeschlagen, benutzt. Die Wahrscheinlichkeit, mit der ein Agent seine Intention einhält (ein Wert von dem die Autoren nicht angeben, wie er berechnet werden kann) wird anhand der hier definierten und voneinander unabhängigen Werte für Altruismus und Ehrlichkeit berechnet. Zum anderen wird die mit dieser Diplomarbeit vorgestellte Erweiterung dieses Vertrauensmodells um eine Kommunikationskomponente beschrieben (Abschnitt 4.1.3). In dieser Komponente wird das Vertrauen und die Bewertung von Aussagen von Agenten *(Zeugen)* über andere Agenten *(Zielagenten)* evaluiert. Der Agent, der diese Komponente nutzt, um Zeugenaussagen zu bewerten, wird der Einfachheit halber *Entscheider* genannt.

4.1.1. Altruismus

Altruismus läßt sich nur in einer Umgebung definieren, in der Aktionen von Agenten möglich sind, die den Zustand anderer Agenten beeinflussen. In dieser Arbeit wählen wir als Grundlage für solche Aktionen das Gefangenendilemma, das wir schon in Abschnitt 2.4 beschrieben haben. Es sei an dieser Stelle noch einmal definiert, in

dem wir gebräuchliche Werte in die Ergebnismatrix aus Abbildung 3 einsetzen (diese Werte wurden z.B. von Bazzan et al. (1997) und Straffin (1996) verwendet).

Definition 5: Gefangenendilemma nach Axelrod (1984)

In einem *Gefangenendilemma* spielen zwei Agenten in einem abstrakten Spiel mit nur zwei Handlungsoptionen, nämlich *Kooperation* (bzw. *C* für *Cooperation*) oder *Verrat* (bzw. *D* für *Defection*). Beide spielen simultan, also ohne Wissen über den Zug des Mitspielers und ohne die Möglichkeit, mit diesem über ihre Züge zu kommunizieren. Je nach dem eigenen und dem gegnerischen Zug erhalten beide Spielpunkte zugewiesen. Diese Spielpunkte müssen bestimmte Relationen erfüllen, wir wählen hier die Spielergebnisse wie sie von Axelrod verwendet wurden (siehe Abbildung 9).

B / A	C	D
C	3 / 3	5 / 0
D	0 / 5	1 / 1

Abbildung 9: Ergebnismatrix des Gefangenendilemmas (nach Axelrod)

Definition 6: Iteriertes Gefangenendilemma

Beim *iterierten Gefangenendilemma* (Axelrod, 1984) handelt es sich um das wiederholte Spielen des Gefangenendilemmas, bei dem die Spieler nicht wissen, wieviel Runden sie spielen werden.

Aus den Regeln des Gefangenendilemmas und der Ergebnismatrix lassen sich eine Reihe von Spielverhalten entwickeln (siehe (Axelrod, 1984), (Delahaye und Mathieu, 1998)). Ein spezielles Verhalten ist absoluter Kooperationswille (Altruismus).

Definition 7: Altruismus, Egoismus, Maß und Modell des Altruismus

Betrachten wir das Verhalten eines Agenten Q gegenüber einem Agenten R im Gefangenendilemma. Als *Altruismus* sei das Verhalten definiert, kooperationswillig zu spielen, auch dann, wenn das Modell von Q über R aussagt, daß R möglicherweise Verrat spielen wird. Ein Handeln gegen den Altruismus wird als *Egoismus* bezeichnet. Das *Maß des Altruismus* eines Agenten Q sei durch die (objektive) Wahrscheinlichkeit definiert, mit der er sich altruistisch verhält:

$$A(Q) = \frac{Anzahl\_Interaktionen(altruistisch)}{Anzahl\_Interaktionen(insgesamt)}$$

Dabei ist die Ereignismenge Ω={„spielt altruistisch“, „spielt nicht altruistisch“}. Die Wahrscheinlichkeit mit der das Ereignis „spielt altruistisch“ eintritt, wird von jedem anderen Agenten X mit Hilfe des von ihm beobachteten bisherigen Verhaltens von Q approximiert. Um die Notation zu vereinfachen benutzen wir

für das *Modell des Altruismus* die selbe Schreibweise wie oben, fügen aber den Name des approximierenden Agenten als Index hinzu:

$$A_X(Q) = \frac{Anzahl\_Interaktionen(altruistisch)}{Anzahl\_Interaktionen(insgesamt)}.$$

4.1.2. Ehrlichkeit

Um den Anforderungen aus Kapitel 3 zu genügen, muß das Gefangenendilemma den Agenten ermöglichen, ihren Spielpartnern mitzuteilen, was sie spielen wollen. Der potentielle Spielpartner muß ebenfalls ankündigen können, was er vorhat zu spielen. Für beide darf diese Ankündigung nicht verpflichtend sein, denn nur dadurch haben beide die Möglichkeit den anderen zu betrügen. Die dahingehende Veränderung des Gefangenendilemmas, das *Offen Gespielte Gefangenendilemma*, wird in der nächsten Definition vorgestellt.

Definition 8: Offen Gespieltes Gefangenendilemma (OGGD)

Folgende Erweiterung des Gefangenendilemmas sei als *offen gespielt* bezeichnet: Betrachten wir die beiden Spieler *A* und *B*. Gemäß dem Protokoll für die Ausführung des Spiels eröffnet Agent *B* die Kommunikation im Spiel, indem er an *A* eine Matrix schickt (siehe Abbildung 10). In dieser Matrix gibt *B* an, welchen Zug *Z* er macht, wenn *A* sich für *C* bzw. *D* entscheidet. Dabei stehen *C* für Kooperation und *D* für Verrat. Danach antwortet *A* welchen Zug er aufgrund dieser Aussage machen wird. Bis zu diesem Zeitpunkt waren alle Angaben nicht verpflichtend, das heißt es ist den Agenten möglich ihren Spielpartner über ihren wahren Zug zu täuschen. Im nächsten Schritt veröffentlichen die Agenten ihren verbindlichen Spielzug gleichzeitig.

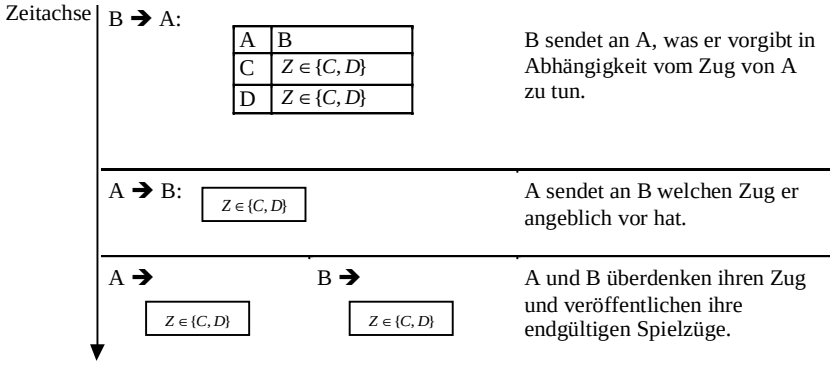

Abbildung 10: Protokoll für den Ablauf des *Offen Gespielten Gefangenendilemmas*

Im Gefangenendilemma gibt es per Definition keine Kommunikation zwischen den beiden Spielern. Daher können die Agenten auch nicht ehrlich oder unehrlich über ihre Züge sein. Diese Schwäche behebt das *Offen Gespielte Gefangenendilemma*.

Definition 9: Ehrlichkeit

Agenten verhalten sich ehrlich, wenn sie das, was sie ankündigen, zu einem zukünftigen Zeitpunkt auch wirklich tun. *Ehrlichkeit* ist das *Normverhalten* in der künstlichen Gesellschaft.

Definition 10: Ehrlichkeit bezüglich Intentionen, Maß und Modell der Ehrlichkeit

Betrachten wir die Handlungen und Intentionen eines Agenten *Q*. Das *Maß der Ehrlichkeit* von *Q* sei definiert als die Wahrscheinlichkeit, mit der sich *Q* gemäß der *Norm* verhält.

$$E(Q) = \frac{Anzahl\_Interaktionen(angekündigt \ldots ausgeführt)}{Anzahl\_Interaktionen(insgesamt)}.$$

Dabei ist die Ereignismenge Ω = {„*Q spielt ehrlich*“, „*Q spielt nicht ehrlich*“}. Die Wahrscheinlichkeit, mit der das Ereignis „spielt ehrlich“ eintritt, wird von jedem anderen Agenten *X* mit Hilfe des von ihm bisher beobachteten Verhaltens von *Q* subjektiv approximiert. Um die Notation zu vereinfachen benutzen wir für das *Modell der Ehrlichkeit* die selbe Schreibweise wie oben, fügen aber den Name des approximierenden Agenten als Index hinzu:

$$E_X(Q) = \frac{Anzahl\_Interaktionen(angekündigt \ldots ausgeführt)}{Anzahl\_Interaktionen(insgesamt)}.$$

Im weiteren soll den Agenten die Möglichkeit gegeben werden, Wissen über das Verhalten von Agenten auszutauschen. Dabei soll auch modelliert werden, daß sie den Empfänger der Daten über ihr eigentliches Wissen täuschen können. Dieses Wissen tauschen die Agenten in Form von Beobachtungen aus. Daher wird nun definiert, was im Offen Gespielten Gefangenendilemma eine Beobachtung ist.

Definition 11: Beobachtung, Menge von Beobachtungen

Eine *Beobachtung* sei ein Tripel (*r*, *alt*, *ehrl)*. Dabei bezeichnet *r* die Spielrunde, in der die Beobachtung gemacht wurde. *alt* steht für eine Variable die `wahr` oder `falsch` ist, je nachdem ob der Agent in dieser Spielrunde altruistisch gehandelt hat oder nicht. Außerdem kann sie auch ε sein, nämlich dann, wenn keine Aussage über diese Spielrunde gemacht wird. Für die Variable *ehrl* gilt analoges. Sie beschreibt, ob der Agent in der betreffenden Runde ehrlich war oder nicht, beziehungsweise daß darüber keine Aussage gemacht wird. Eine *Menge von m Beobachtungen* ist ein Tupel der Länge *m*. Elemente des *m*-Tupels sind Beobachtungstripel.

Nun kann auch ein Versuch unternommen werden, den Begriff der Ehrlichkeit aus Definition 9 auf das Kommunizieren von Beobachtungen zu erweitern:

Definition 12: Ehrlichkeit bezüglich der Kommunikation von Beobachtungen

Ehrlichkeit bezüglich der Kommunikation von Beobachtungen besteht darin, daß ein Agent, wenn er über das Verhalten eines anderen befragt wird, alle Daten wahrheitsgemäß übermittelt. Wenn man so will, soll er *offen* und *ehrlich* sein. *Ehrlichkeit bezüglich der Kommunikation von Beobachtungen* liegt also insbesondere dann vor, wenn ein Verstoß gegen diese Norm dazu führen würde, daß der Empfänger der Daten einen anderen Agenten als besseren Spielpartner einstuft als dem Informanten recht ist.

Diese Definition ist jedoch nicht präzise genug, um implementiert zu werden. Daher wird das Problem noch einmal aus einer anderen Perspektive betrachtet. Es soll das Gegenteil betrachtet werden: Was bedeutet es zu *betrügen*? Dazu betrachten wir, was die Motive zum Lügen in den Szenarien sind.

Jeder Agent hat das Ziel, sich möglichst vielen Agenten als Spielpartner anzupreisen, um sich dann den für ihn besten aussuchen zu können. Wird er von einem Agent befragt, der gerade dabei ist, sich einen Spielpartner zu suchen, wird er natürlich versuchen, möglichst alle anderen Agenten in einem schlechten Licht erscheinen zu lassen. Er hat also kein Interesse daran, das Schlechte was er über andere weiß, zu verheimlichen. Was könnte er stattdessen tun, um andere in einem schlechten Licht erscheinen zu lassen? Er könnte zum einen negative Beobachtungen erfinden, oder positive Beobachtungen verheimlichen (zu Vertrauen und Kommunikation siehe auch (Zand, 1977) und den Abschnitt 2.3.1 zum Thema *Mißtrauen*). Damit der Zeuge Spielbeobachtungen, in denen der Zielagent etwas Negatives getan hat, erfinden könnte, müßte er auch die Spielrunde, in der dies angeblich geschah, angeben können. Dies würde dann aber die Gefahr bergen, daß der Entscheider das entsprechende Spiel selbst beobachtet hat. Der könnte dann sofort Rückschlüsse über den Zeugen ziehen. Um nicht durch allzu offensichtlichen Betrug entlarvt zu werden, bleibt dem Zeugen daher nur die Möglichkeit positive Beobachtungen über den Agenten zu verheimlichen. Daran hindert ihn allein die Norm der Ehrlichkeit. Dieser Norm gehorchen die Agenten mit einer gewissen Wahrscheinlichkeit (siehe Definition 9). Daraus folgt nun direkt Definition 13:

Definition 13: Betrügen

Sei *Betrügen* eine Abbildung:

$$\textit{Betrügen}: \{\textit{wahr}\}\blacklozenge\ \{\textit{wahr}, \varepsilon\}$$

$$x \mapsto \begin{cases} \textit{wahr} & \text{mit Wahrscheinlichkeit p} \\ \varepsilon & \text{mit Wahrscheinlichkeit 1 - p} \end{cases}$$

Dies ist eine Abbildung, die eine positive Aussage über einen anderen Agenten abbildet auf einer leeren Aussage (dem Verschweigen dieser Information) oder aber der identischen Information. Der Wert der Abbildung wird anhand eines Parameters *p* bestimmt, der angibt, mit welcher Wahrscheinlichkeit die Abbildung welchen Wert annimmt. Dabei sei die Ereignismenge $\Omega = \{$„*positive Aussage wird berichtet*“, „*positive Aussage wird nicht berichtet*“$\}$. Für den Fall des Betrügens bzgl. Beobachtungen von Spielen in denen ein Agent ehrlich war, sei die Ereignismenge $\Omega = \{wahr, \varepsilon\}$. Dabei steht *wahr* für die Aussage, daß ein anderer Agent ehrlich war und ε für keine Aussage. Der Parameter *p* gibt an, wie häufig nicht betrogen wird.

Betrachten wir die Anwendung der in Definition 13 beschriebenen Funktion. Jede Beobachtung, die die Information enthält, daß der Zielagent sich unehrlich verhalten hat, wird direkt weitergeben. Bei jedem Tripel der Form *(r, alt, wahr)* wird beim Betrügen bzgl. der Ehrlichkeit auf das letzte Element die *Betrügen*-Funktion angewendet. Dadurch ergibt sich entweder *(r, alt, wahr)* oder *(r, alt, ε)* als Tripel, das er an den Entscheider weitergibt. Betrügt ein Agent bzgl. beider Attribute (Ehrlichkeit und Altruismus), wird die Funktion für beide Variablen *(alt* und *ehrl)* getrennt aufgerufen.

4.1.3. Vertrauen in Zeugen und ihre Aussagen

Wir betrachten nun die Erweiterung des Vertrauensmodells um eine Komponente für die Berechnung des Vertrauens in Zeugen und ihre Aussagen. Im wesentlichen beruht der Algorithmus darauf, mit Mitteln der Mathematik zu schätzen, wieviel Information verheimlicht wurde und zu versuchen diese zu rekonstruieren. Wir stellen hier Gleichungen vor, die diese Rekonstruktion leisten können.

Wird für *Betrügen* die Definition 13 zugrundegelegt, dann handelt es sich dabei mathematisch gesprochen um ein *Bernoulli-Experiment*. Es entspricht dem Zufallsexperiment des Werfens einer gezinkten Münze. Zeigt die Münze Kopf, so lügt der Agent über die entsprechende Beobachtung, bei Zahl teilt er sie mit. Ein wiederholtes Werfen dieser Münze entspricht einer *Bernoulli-Kette*. Normalerweise wird die Bernoulli-Kette für die Berechnung der Wahrscheinlichkeit benutzt, mit der bei n-maligem Werfen die Münze k-mal Kopf bzw. Zahl zeigt (Berechnung über die Binomialverteilung). Im folgenden stellen wir vor, wie die Bernoulli-Kette benutzt werden kann, um Herauszufinden wie oft höchstwahrscheinlich eine Münze geworfen wurde, wenn nur bekannt ist, daß k-mal das Ergebnis Kopf vorkam. Dies entspricht dem, was der Entscheider über den Zeugen weiß: Er weiß, wie oft dieser als Ergebnis seiner *Betrügen*-Funktion das Ergebnis *wahr* hatte. Mit anderen Worten, der Entscheider kennt die Mindestanzahl der ehrlichen Antworten (diese bezeichnen wir im folgenden mit e), da er davon ausgehen kann, daß der Zeuge keine positiven Beobachtungen über den Zielagenten erfindet. Er weiß aber nicht, wie oft das Ergebnis Verschweigen einer Beobachtung, also ε, war.

Spielrunde	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Betrogen
Zeuge				✓		✓			**X**		**X**	**X**		**X**	?

Abbildung 11: Beispiel für die übermittelten Daten eines Zeugen

Ein Beispiel für eine solche Situation ist in Abbildung 11 gezeigt. Ein „√“ bedeutet in der Abbildung die (mitgeteilte) Beobachtung eines ehrlichen Verhaltens, ein „X“ ist eine Beobachtung eines unehrlichen Verhaltens. Der Zeuge berichtet über zwei ehrliche Spielzüge (in Runde vier und sechs). Damit ist $e = 2$. Es ist nicht bekannt, wie oft er über eine ehrliche Beobachtung geschwiegen hat (k ist unbekannt). Es soll nun beschrieben werden, wie der Entscheider die Anzahl des Betrügens approximieren kann. Danach wird dargestellt, wie mehrere Zeugenaussagen miteinander kombiniert werden können. Es seien folgende Variablen definiert:

Definition 14: Parameter n, k, e, p für die Vertrauensberechnung
Es seien:

- n die Anzahl der Spielrunden, die ein Zeuge über einen Zielagenten beobachtet hat.
- k die Schätzung der Anzahl, wie oft das Ergebnis der *Betrügen*-Funktion ε war, d.h. k ist die Approximation davon, wie oft der Zeuge sich entschlossen hat, während der Kommunikation eine Information über den Zielagenten zu verheimlichen.
- e die Anzahl der ehrlichen Spielrunden des Zielagenten über die der Zeuge berichtet hat. Dies bezeichnet wie oft die *Betrügen*-Funktion *wahr* ergab.

- p eine Schätzung des Entscheiders für den gleichnamigen Parameter der *Betrügen*-Funktion des Zeugen. Dieser Parameter ist entweder das Ergebnis einer Schätzung aufgrund von eigenen Beobachtungen oder aber ein Resultat einer vorhergehenden Anwendung dieses Algorithmus auf den Zeugenagenten.

In der Berechnung wird der Parameter p benutzt. Dies bedeutet, daß zur Bewertung der Aussage über einen Agenten zunächst die Ehrlichkeit eines Zeugen evaluiert werden muß. Dazu müssen entweder eigene Beobachtungen über den Zeugen vorliegen, oder aber eine vorherige Anwendung dieser Berechnung hat einen Wert für dessen p ergeben. Es wurde schon die Formel für die Binomialverteilung angesprochen. Diese gibt an, mit welcher Wahrscheinlichkeit bei n Experimenten k-mal das Ereignis eintritt, dessen Wahrscheinlichkeit q ist:

$$P(X = k) = \frac{n!}{k!(n-k)!} q^k (1-q)^{n-k}. \qquad (9)$$

Dabei steht X für die Wahrscheinlichkeitsgröße „*Wie oft trifft das Ereignis ein*“. Im folgenden ist aber nicht von Interesse, wie oft insgesamt gespielt wurde, sondern nur wie oft betrogen wurde. Wir ersetzen also n durch $k+e$. Außerdem ist nicht die Wahrscheinlichkeit q für das Betrügen bekannt, sondern die Wahrscheinlichkeit p für das „ehrlich sein“. Da $p = 1-q$ ergibt sich für die Binomialverteilung:

$$P(X = k) = \frac{(k+e)!}{k!e!} (1-p)^k p^e. \qquad (10)$$

Unser Ziel ist es, zu berechnen, welches k das Wahrscheinlichste ist. Wir benötigen also die Berechnung des *Erwartungswertes EX* für diese Verteilung. Dabei unterscheiden wir zwei Fälle in Abhängigkeit des übermittelten e.

Fall $e > 0$**:** Sobald der Zeuge über mehr als ein e berichtete, läßt sich der Erwartungswert EX für die Binomialverteilung folgendermaßen berechnen (siehe z.B. Bronstein und Semendjajew (1991)):

$$\begin{aligned} EX &= \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} q^k (1-q)^{n-k} \\ &= nq \end{aligned}$$

Nach Definition 14 ist k gerade so definiert, daß es gleich dem Erwartungswert EX ist. Setzen wir k für EX ein, so erhalten wir:

$$k = nq$$

Wir setzen nun wie oben für n und q unsere Variablen ein: $k+e$ für n, $1-p$ für q. Wir erhalten:

$$k = (k+e)(1-p)$$

$$\Leftrightarrow k = \frac{e}{p} - e. \qquad (11)$$

Damit ist eine Approximation gegeben für die Anzahl der Lügen des Zeugen bei denen er wissentlich Information in seiner Aussage weggelassen hat. Diese

Berechnung des Erwartungswertes gilt aber nur falls $e > 0$. Wenn $e = 0$ ist, dann liefert diese Gleichung keine gewinnbringende Lösung. Das liegt daran, daß keine Minimale Datenmenge da ist von der nach oben interpoliert werden kann. Daher muß ein anderer Weg beschritten werden. Der Fall $e = 0$ Fall tritt dann ein, wenn ein Zeuge sehr häufig lügt (p sehr groß ist), oder er selbst noch sehr wenig positive Informationen hat (n sehr klein).

Fall e = **0 :** Betrachtet man die für den obigen Erwartungswert zugrundeliegende Gleichung (10) für die Binomialverteilung, und setzt man für e den Wert *0* ein, so erhält man:

$$P(T = k) = \frac{k!}{k!}(1-p)^k$$
$$= (1-p)^k$$

Wie im Fall $e > 0$ wird nun der Erwartungswert EX dieser Funktion berechnet. Nach gängigem Verfahren (siehe z.B. (Bronstein und Semendjajew 1991)) geschieht dies durch Integration, d.h. der Berechnung der Fläche unter der Kurve. Ist diese bestimmt, wird die Stelle auf der x-Achse berechnet, an der das Maß der Fläche rechts von dieser Stelle gleich dem Maß links davon ist.

Wenn wir dieses Verfahren auf den Term aus der Binomialverteilung anwenden bedeutet dies:

$$\frac{1}{2}\int_0^{\infty}(1-p)^k dk = \int_0^{EX}(1-p)^k dk. \qquad (12)$$

Die Integration in dieser Gleichung wird gelöst durch Substitution. Betrachten wir daher zunächst nur das unbestimmte Integral

$$\int (1-p)^k dk.$$

Wir substituieren $t = (1-p)^k$.

Damit ist $k = \frac{\ln t}{\ln(1-p)}$ und $dk = \frac{1}{\ln(1-p)}\frac{1}{t}dt$. Wir erhalten also:

$$\int (1-p)^k dk = \int \frac{1}{\ln(1-p)} d\,\mathrm{t}$$
$$= t\frac{1}{\ln(1-p)} + c.$$

Die Rücksubstitution $t = (1-p)^k$ ergibt:

$$\int (1-p)^k dk = (1-p)^k \frac{1}{\ln(1-p)} + c. \qquad (13)$$

Dies kann nun benutzt werden um den Erwartungswert EX zu bestimmen. Wenn wir (13) benutzen um beide Seiten von (12) auszurechnen, erhalten wir folgende Gleichung:

$$\frac{1}{2}\lim_{a\to\infty}\left[(1-p)^k\frac{1}{\ln(1-p)}\right]_0^a=\left[(1-p)^k\frac{1}{\ln(1-p)}\right]_0^{EX}$$

$$\Leftrightarrow\frac{1}{2}\lim_{a\to\infty}\left((1-p)^a\frac{1}{\ln(1-p)}-\frac{1}{\ln(1-p)}\right)=(1-p)^{EX}\frac{1}{\ln(1-p)}-\frac{1}{\ln(1-p)}$$

Da der Grenzwert für $0\le p\le 1$ und für $a\to\infty$ aber gegen 0 geht, folgt:

$$-\frac{1}{2}\frac{1}{\ln(1-p)}=(1-p)^{EX}\frac{1}{\ln(1-p)}-\frac{1}{\ln(1-p)}$$

$$\Leftrightarrow\left(-\frac{1}{2}+1\right)\frac{1}{\ln(1-p)}=(1-p)^{EX}\frac{1}{\ln(1-p)}$$

$$\Leftrightarrow\quad\frac{1}{2}\frac{1}{\ln(1-p)}=(1-p)^{EX}\frac{1}{\ln(1-p)}$$

$$\Leftrightarrow\quad\frac{1}{2}=(1-p)^{EX}$$

$$\Leftrightarrow\quad EX=\frac{\ln 1/2}{\ln(1-p)}$$

Wie im ersten Fall setzen wir gemäß Definition 14 $k = EX$. Insgesamt ergibt sich also aus Formel (12), daß der Erwartungswert für die Anzahl der Lügen k von der Wahrscheinlichkeit p, mit der der Zeuge ehrlich ist, folgendermaßen abhängt:

$$k=\frac{\ln 1/2}{\ln(1-p)}. \qquad (14)$$

Es zeigt sich, daß wir k (nach Anwendung einiger mathematischer Regeln) überraschend einfach abschätzen können. Jetzt kann in jedem Fall für jeden Zeugen, für den eine Approximation seiner Ehrlichkeit p berechnet wurde, approximiert werden, wie oft er betrogen hat, als er Beobachtungen über den Zielagenten mitgeteilt hat. Für Abbildung 11 bedeutet dies, daß der Entscheider einen Wert für das Feld „Betrogen“ einsetzen kann. Diese Information wird später im nächsten Kapitel benötigt, um die Informationen über das Betrügen von mehreren Zeugen bezüglich eines Agenten zu evaluieren.

4.1.4. Vertrauen in Kooperationspartner

Mit diesen Definitionen sind die Normen der künstlichen Gesellschaft und prinzipielle Verhaltensweisen definiert. Daraus ergeben sich drei mögliche Verhaltensweisen in jeder Runde des Offen Gespielten Gefangenendilemmas (siehe Abbildung 12). Jeder Agent entscheidet sich ob er sich altruistisch oder egoistisch verhalten soll.

Dabei soll zunächst außer acht gelassen werden, wie er zu dieser Entscheidung kommt.

Hat er sich für *egoistisch* entschieden bleiben ihm zwei Handlungsoptionen. Er kann dies zugeben, also *ehrlich* sein und sagen, er wird Verrat spielen, oder aber er kann *Betrügen* und behaupten, er wird im nächsten Zug altruistisch spielen. Hat der Agent sich für einen altruistischen Spielzug entschieden, so macht es für ihn keinen Sinn zu behaupten, er wird seinen Mitspieler verraten, da er dadurch sein Spielergebnis nicht verbessern kann. Sein Spielpartner wird auf jeden Fall ein gleiches oder schlechteres Ergebnis erwarten, wenn der Verrat als Handlungsoption angegeben wurde (siehe Abbildung 9). Deshalb entfällt dieser vierte Ast im Entscheidungsbaum.

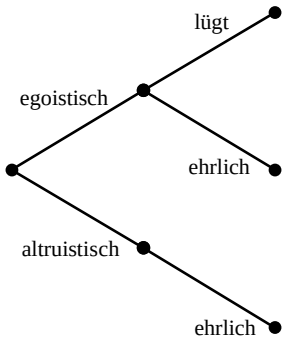

Abbildung 12: Entscheidungsbaum für das OGGD

Betrachten wir nun die Situation, in der ein Agent X sich befindet, wenn er die Aussage erhält, daß ein Agent Q kooperativ mit ihm spielen will. Agent X muß auf jeden Fall vorsichtig sein, denn Agent Q könnte gelogen haben und eigentlich vorhaben, X auszubeuten. Eine Analyse des Entscheidungsbaumes ergibt, daß Q zwei mögliche Pfade gegangen sein kann, um zu der Ankündigung *Kooperation* gekommen zu sein (Abbildung 13). Wäre Q egoistisch und ehrlich, so hätte er nicht *Kooperation*, sondern *Verrat* angekündigt.

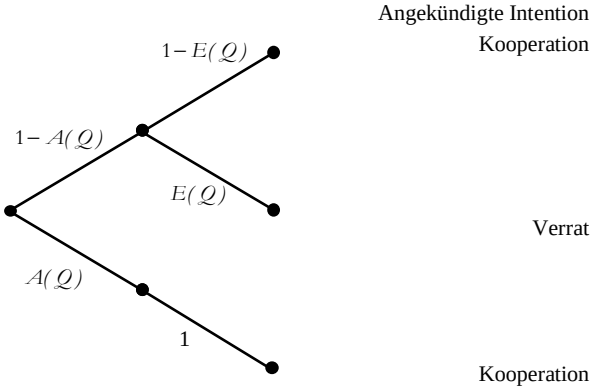

Abbildung 13: Berechnung der Vertrauenswürdigkeit

Es ist nun im Interesse von X die Wahrscheinlichkeit zu bestimmen, mit der Q wirklich altruistisch spielen will. Dies entspricht dem im Bild untersten Pfad. Voraus-

setzung für diese Bestimmung ist, daß X ein Modell davon hat, mit welcher Wahrscheinlichkeit der Agent Q an jedem der Entscheidungspunkte einen der beiden rechten Äste gewählt hat. Die Kanten in Abbildung 13 sind mit den Werten annotiert, die X für eine Einschätzung des Verhaltens von Q approximieren muß. Diese Werte werden in der folgenden Definition zusammengefaßt.

Definition 15: Modell eines Agenten

Seien $A_X(Q)$ und $E_X(Q)$ definiert wie in Definition 7 und Definition 10. Das Paar $(A_X(Q), E_X(Q))$ bezeichnen wir als *Modell* des Agenten X von Agent Q.

Soll ein Agent X entscheiden, wie er das Verhalten eines Agenten Q einschätzt, so tut er dies also aufgrund seines Modells von Q. Sind die Wahrscheinlichkeiten für die einzelnen Entscheidungspunkte bekannt, so läßt sich auch berechnen, wie hoch die Wahrscheinlichkeit für den unteren Pfad im Entscheidungsbaum ist. Da X nicht weiß, wie die Wahrscheinlichkeiten des Verhaltens von Q sind, versucht er aufgrund von Beobachtungen des Verhaltens die Werte sein Modell anzunähern. Das bedeutet, daß er die Werte für $A_X(Q)$ (der Approximation des Altruismus von Q) und $E_X(Q)$ (der Approximation der Ehrlichkeit von Q) bestimmt. Sind diese Werte bestimmt, so läßt sich annähern, wie hoch die Wahrscheinlichkeit dafür ist, daß Q den unteren Entscheidungsast gewählt hat. Diese Wahrscheinlichkeit W entspricht nämlich dem Anteil des unteren Astes an der Summe der Wahrscheinlichkeiten beider Äste die in einem Angebot der Kooperation münden:

$$W = \frac{P(Q\ altruistisch)}{P(Q\ altruistisch) + P(Q\ egoistisch \cap Q\ lügt)}$$
$$= \frac{P(Q\ altruistisch)}{P(Q\ altruistisch) + P(Q\ egoistisch)P(Q\ Lügt)}.$$

Dieser Wert gibt wieder, wie hoch die Wahrscheinlichkeit ist, daß der Agent sein Wort halten wird. Dies wird der folgenden Definition als Vertrauenswürdigkeit definiert.

Definition 16: Vertrauenswürdigkeit eines Kooperationsangebots

Unter der Voraussetzung, daß ein Agent Q einem Agenten X kooperatives Verhalten zugesichert hat, ist die *Vertrauenswürdigkeit* von Q definiert als die Abschätzung der Wahrscheinlichkeit W, mit der sich auch wirklich kooperativ verhalten wird.

Dies soll noch an einem Beispiel verdeutlicht werden. Gegeben sei ein Agent Q, dessen Verhalten charakterisiert werden kann mit der Wahrscheinlichkeit von $2/3$ für seinen Altruismus und $3/4$ für seine Ehrlichkeit. Damit ergibt sich für ihn ein W von:

$$W = \frac{\frac{2}{3}}{\frac{2}{3} + \frac{1}{3}\frac{1}{4}} = \frac{\frac{8}{12}}{\frac{8+1}{12}} = \frac{8}{9} \approx 0{,}889.$$

Hätte man nur den Altruismus oder die Ehrlichkeit des Agenten betrachtet, hätte man die Chancen dafür, daß er wirklich kooperiert auf $0{,}\overline{6}$ bzw. auf 0,75 geschätzt. Nur wenn beide Faktoren, über die man ja bereits Wissen angesammelt hat,

betrachtet werden, kommt man auf die wesentlich exaktere Einschätzung des Verhaltens von *Q*.

Dies spielt insbesondere dann eine Rolle, wenn sich mehrere Agenten anbieten, die einen ähnlichen Ehrlichkeitswert haben, aber unterschiedliche Altruismuswerte besitzen. Betrachten wir einen zweiten Agenten *R,* der sich neben *Q* als Spielpartner anbietet. Sein Wert für Ehrlichkeit sei mit 0,8 etwas höher gewählt, als der entsprechende Wert von *Q* (0,75). Sein Altruismuswert sei gleich $0,\overline{3}$ (statt $0,\overline{6}$). Für einen Entscheider, der nur die Ehrlichkeit betrachtet spielt das aber keine Rolle, er würde sich auf den um fast 10 Prozent höheren Ehrlichkeitswert verlassen. Ein Entscheider, der jedoch die Vertrauenswürdigkeit berechnet erhält folgendes W:

$$W = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{2}{3}\frac{2}{10}} = \frac{\frac{5}{15}}{\frac{5+2}{15}} = \frac{5}{7} \approx 0,715.$$

Ein Agent, der nur die Ehrlichkeit betrachtet, hätte also einen Agenten als Spielpartner vorgezogen, der nur mit einer Wahrscheinlichkeit von 0,715 sein Versprechen hält und einem Spielpartner abgesagt, der mit einer Wahrscheinlichkeit von 0,889 kooperiert hätte. Damit hätte er seine Chancen auf eine gelungene Kooperation um mehr als 15 Prozent verschlechtert. Den selben Fehler begehen Entscheider, die zwar den Altruismus betrachten, aber die den Wert der Ehrlichkeit vernachlässigen. Auch hier ein Beispiel: Sei *S* ein Agent mit einem im Vergleich höheren Altruismuswert 0,75 (*Q* hat eine Wert von 0,667) und dem Ehrlichkeitswert 0,25 (statt 0,75). Damit ergibt sich für *W*:

$$W = \frac{\frac{3}{4}}{\frac{3}{4} + \frac{1}{4}\frac{3}{4}} = \frac{\frac{12}{16}}{\frac{12+3}{16}} = \frac{12}{15} = 0,8.$$

Zwar war der Altruismus von *S* um 10 Prozent höher, aber die Wahrscheinlichkeit, daß er den Entscheider trotzdem betrügen wird, liegt immer noch bei 20 Prozent. Wählt der Entscheider den Agenten *Q* als Spielpartner, so hat dieser zwar nur einen Altruismus von 0,667, aber die Wahrscheinlichkeit, daß er sein Angebot einhält, liegt bei 0,889 und damit um mehr als 10 Prozent über der von *S*.

Zusammenfassend läßt sich sagen, daß die Analyse von Ehrlichkeit oder Altruismus auf lange Sicht im Gefangenendilemma nicht ausreicht. Erst die Analyse von Vertrauen, nämlich beider Faktoren, führt zu einer wirklich guten Approximation des Verhaltens der anderen Agenten. Verfügt ein Agent *X* über die oben genannten Approximationsfunktionen für die Ehrlichkeit und den Altruismus eines Agenten *Q*, erhält er durch Einsetzen folgenden Wert für *W*:

$$W = \frac{A_X(Q)}{A_X(Q) + (1 - A_X(Q))(1 - E_X(Q))}.$$

Dieser Wert wird in der folgenden Definition benutzt, um das Verhalten eines Agenten einzuschätzen.

Definition 17: Maß der Vertrauenswürdigkeit

Das *Maß der Vertrauenswürdigkeit* eines Agenten Q aus der Sicht eines Agent X ($V_X(Q)$) unter der Voraussetzung, daß Q Kooperation angeboten hat, sei definiert als die Wahrscheinlichkeit W, mit der X davon ausgeht, daß Q sein Wort halten wird. Es gilt also:

$$V_X(Q) = \frac{A_X(Q)}{A_X(Q) + (1 - A_X(Q))(1 - E_X(Q))}. \quad (15)$$

Damit kann nun auch eine Definition von Vertrauen angegeben werden, die sowohl eine Begründung in den Definitionen der Psychologie findet, als auch in einem mathematischen Kontext formalisierbar ist.

Definition 18: Vertrauen, Maß des Vertrauens

Kommt ein Agent X nach der Evaluation der *Vertrauenswürdigkeit* eines Agenten Q zu dem Schluß, sich auf dessen Angebot der Kooperation einzulassen und mit ihm zu spielen, so sagen wir: X *vertraut* Q bezüglich eines Kooperationsangebotes.

Das *Maß* dieses *Vertrauens* entspricht dem aus subjektiver Sicht ermittelten *Maß der Vertrauenswürdigkeit.*

Hierbei wird explizit zwischen *Vertrauen* und *Vertrauenswürdigkeit* unterschieden. Der Unterschied liegt in der Perspektive und wurde der sauberen Trennung der Begriffe wegen beschrieben. Vertrauenswürdigkeit bezieht sich auf eine Einschätzung (*belief*) eines Agenten. Vertrauen ist jedoch, wie in der Definition von Deutsch (siehe Definition 1), eine innere Haltung oder ein Zustand des Agenten bezüglich einer bestimmten *Interaktion mit einem Agenten*. Theoretisch ist auch denkbar, daß sich beide unterscheiden, etwa wenn verschiedene Interaktionen verschiedene Wichtigkeit haben. Dann kann die Vertrauenswürdigkeit eines anderen Agenten sehr niedrig sein, das Vertrauen kann aber trotzdem bezüglich einer einzelnen, aber unwichtigen Interaktion sehr hoch sein.

Diese Evaluation entspricht dem Modell von Castelfranchi und Falcone mit einigen Modifikationen. Was sie in ihrem quantitativen Modell als *Trust* bezeichnen, wird hier Vertrauenswürdigkeit genannt. Bei der Berechnung des *DoT (Degree of Trust)* werden fünf Werte miteinander multipliziert. Die beiden Autoren schlagen eine Semantik dieser Werte vor, geben jedoch keinen Hinweis zur Berechnung. Wir haben mit Definition 18 eine Berechnung für den Wert, den sie *Intention* nennen angegeben. Die Modelle sind ansonsten identisch, wenn man die anderen Faktoren in der Formel (4) von Castelfranchi und Falcone auf *1* setzt

4.2. Experimentalumgebung

Wir stellen nun eine Experimentalumgebung vor, in der Agenten autonom agieren können (Abschnitt 4.2.1). Sie spielen eine speziell für diese Problemstellung entworfene Abwandlung des *Gefangenendilemma*, nämlich das *Offen Gespielte Gefangenen-dilemma* (siehe auch Definition 8), das wir um eine Phase der Partnerauswahl erweitert haben. Durch die Möglichkeit andere Agenten bei ihren Spielen zu beobachten, sind die Agenten in der Lage, eine qualifizierte Entscheidung darüber zu treffen, ob es ihnen dient mit einem Agenten zu kooperieren oder nicht. Diese Entscheidung

beruht auf eigenen Beobachtungen, der Benutzung der Beobachtungen anderer und der Einschätzung der Vertrauenswürdigkeit dieser Zeugen. In dieser Experimentalumgebung hat jeder Agent die Möglichkeit, sich seinen Spiel- und Kooperationspartner auszusuchen. Dieser versetzt ihn in die Lage, vorteilhafte Kooperationen einzugehen oder bestimmte Agenten zu meiden. Die gemiedenen Agenten haben, wenn sie keine Spielpartner finden, keine Möglichkeit, Ressourcen für sich zu erhalten. In dieser Experimentalumgebung können Agentenstrategien für die in Abschnitt 3.3 beschriebenen Szenarien getestet und evaluiert werden. Dabei gehen wir von weichen Persönlichkeitsprofilen aus, die einen kontinuierlichen Übergang zwischen egoistisch und altruistisch modellieren. Eine Rechtfertigung dafür, daß dieses Testbett eine realistische Modellierung ist, findet sich in 4.2.2. Die dazugehörige Implementierung wird im nächsten Kapitel beschrieben.

4.2.1. Offen Gespieltes Gefangenendilemma mit Partnerauswahl

Im Offen Gespielten Gefangenendilemma mit Partnerauswahl erhält jeder der Spieler zu Beginn zwanzig Punkte. Es ist definiert, wieviel Agenten sich jeweils in einer Nachbarschaft befinden. Von dieser Nachbarschaft hängt es ab, wieviel andere Spieler ein Agent beobachten kann. Der Ablauf des Spiels gliedert sich in fünf Phasen:

1. Jeder Spieler zahlt einen Punkt dafür, daß er in dieser Runde mitspielen darf.
2. Es werden die Spielpaare für diese Runde ermittelt. Die Agenten können frei entscheiden, ob sie mit einem Agenten zusammen spielen wollen oder nicht.
3. Die Spielpartner, die beide miteinander spielen wollen, verpflichten sich simultan zu einem Spielzug
4. Jeder Agent beobachtet das Spielverhalten seiner Nachbarn.
5. Auszahlung der Gewinne.

Die hier im Überblick gezeigten Phasen werden nun im Detail erklärt.

Phase 1 Einzahlung

Jeder Agent, der sich entscheidet mitzuspielen, zahlt einen Punkt für dieses Spiel. Nur Agenten, die bezahlt haben, nehmen an dem Spiel teil. Ein Spieler, der keine Punkte mehr hat, scheidet aus dem Spiel aus.

Phase 2 Partnerauswahl

Um eine Annäherung an die Problemstellung zu erreichen, wurde das Gefangenendilemma insofern erweitert, daß die Agenten die Möglichkeit haben auf die Wahl ihres Spielpartners Einfluß zu nehmen. Die Ermittlung der Spielpaare erfolgt nach einem dem *contract net protocol* ähnlichen Protokoll, das so oft hintereinander ausgeführt wird, bis alle Agenten die Chance gehabt haben, einen Spielpartner zu finden (siehe Abbildung 14).

Zunächst werden die Agenten in einer zufälligen Reihenfolge in einer Liste angeordnet. Da die Gruppe der Agenten potentiell heterogen ist, muß sichergestellt

sein, daß die Reihenfolge der Agenten keinen Einfluß auf das Ergebnis des Experimentes nimmt.[1]

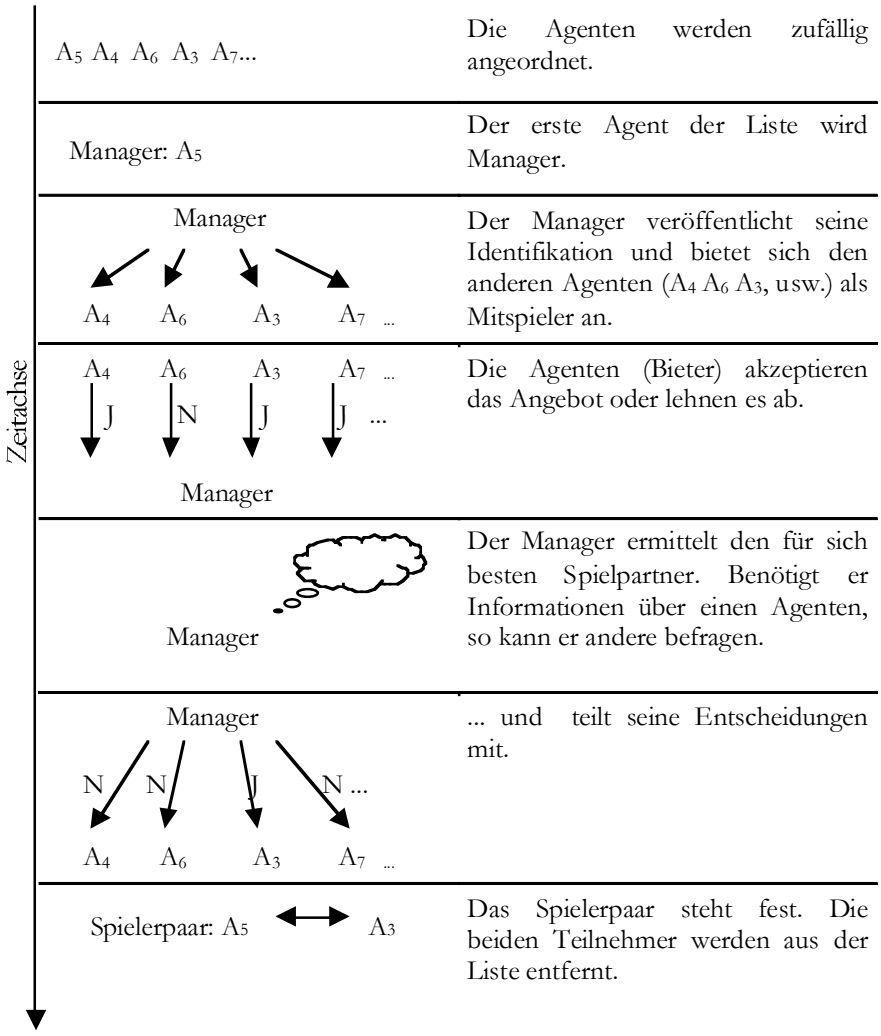

Abbildung 14: Die Agenten wählen einen Spielpartner aus

In jeder Runde dieses Verfahrens wird ein Agent aus der Liste ausgewählt. Gemäß dem Standard-CNP-Verfahren wird er im folgenden *Manager* genannt. Der Manager ist derjenige Agent, dem es möglich ist, sich als Spielpartner anzubieten. Er identifiziert sich gegenüber den anderen Agenten und kündigt gemäß dem ersten Teil des *Offen gespielten Gefangenendilemmas* an, wie er spielen wird (siehe auch Definition 8). Die anderen Agenten haben nun die Möglichkeit sich ihm wiederum als Spielpartner anzubieten, oder dies abzulehnen und ihm ihre Spielzüge anzukündigen. Sobald alle Agenten sich geäußert haben, hat der Manager Gelegenheit sich zu entscheiden, mit wem er spielen will. Den Agenten ist daran gelegen mit einem anderen Altruisten zu spielen, da ein Spiel mit einem Altruisten gemäß der Ergebnismatrix in jedem Fall einen höheren Gewinn sichert. In den Szenarien der möglichen Anwendungen ist es

[1] Dieser Effekt wird in der Analyse weiter dadurch eliminiert, daß eine große Anzahl von Experimenten durchgeführt wird und deren Ergebnisse gemittelt werden.

so, daß der angekündigte Spielzug ein wichtiges Kriterium für die Auswahl des Spielpartners ist. Daher findet in diesem der erste Teil des Offen Gespielten Gefangenendilemmas schon in dieser Phase statt. Die Agenten kündigen ihren Zug gleichzeitig mit ihrer Identifikation im Spielangebot an.

Es ist jedoch Teil des Experimentes, daß der Agent nicht weiß, wie sich seine potentiellen Spielpartner verhalten („*The world does not come labelled*" (Edelmann, 1987)). Um dieses Problem zu lösen hat der Manager in dieser Spielphase die Gelegenheit mit allen Agenten zu kommunizieren, um mehr über seine eventuellen Spielpartner zu erfahren. Bei dieser Kommunikation sind Anfrage und Antwort strikt geregelt: Der Manager richtet eine Anfrage an einen Agenten, z.B. D. Inhalt der Anfrage ist der Name eines potentiellen Spielpartners Q. Agent D antwortet daraufhin mit seinen Beobachtungen über Q. Diese müssen nicht seinen wahren Beobachtungen entsprechen. Die Anzahl der erlaubten Anfragen wird durch die Implementierung eingeschränkt und zählt zu den untersuchten Variablen (siehe auch Abschnitt 6.2). Mit Hilfe einer geeigneten Datenstruktur wertet der Manager die eigenen Beobachtungen und die von ihm durch Kommunikation gesammelte Information aus (für Details über die Datenstruktur, siehe Abschnitt 0). Hat sich ein Spielpaar gefunden, so werden beide Agenten aus der Liste ausgetragen.

Findet ein Agent keinen Spielpartner, so muß er für diese Runde aussetzen. Er erhält seinen Einsatz zurück. Die Entscheidung des Miteinanderspielens beruht bedingt durch das Protokoll auf Gegenseitigkeit. Das Verfahren ist iterativ. Es wird solange durchlaufen, bis alle Agenten die Gelegenheit hatten, sich für einen Spielpartner zu entscheiden. War ein Agent einmal Manager, wird er es nicht ein zweites Mal. Er kann aber als Bieter dem Spielangebot eines anderen Agenten zustimmen und hat damit weitere Gelegenheiten noch in dieser Runde an einer Interaktion teilzunehmen.

Phase 3 Der verpflichtende Spielzug

Diese Phase erfordert von jedem Agenten, der einen Spielpartner gefunden hat, einen verpflichtenden Spielzug. Dieser erfolgt simultan mit dem des Spielpartners. Ein Reagieren auf dessen Zug ist also nicht möglich. Der Agent kann jedoch auf die Ankündigung seines Spielpartners eingehen. Eventuell bricht er damit aber seine eigene Ankündigung. Diese Phase stellt den zweiten Teil des in Definition 8 vorgestellten *Offen Gespielten Gefangenendilemmas* dar. Der Spielzug in dieser Phase ist verbindlich für die Verteilung der Punkte.

Phase 4 Beobachtung der Spielergebnisse

Alle Agenten haben nun die Möglichkeit, sowohl die als verbindlich angegebenen Züge, als auch die angekündigten Züge ihrer Nachbarn zu erfahren. In der Implementierung kann angegeben werden, wieviel Agenten zu einer solchen Nachbarschaft gehören. Die Größe dieser Nachbarschaft ist eine weitere untersuchte Variable in der Evaluation des Ansatzes (siehe Kapitel 6). Damit hat jeder Agent nun die Möglichkeit, Berechnungen über die Ehrlichkeit und das Verhalten (egoistisch vs. altruistisch) anzustellen. Diese Daten kann er dann bei der nächsten Partnerauswahl und bei der nächsten Kooperationsentscheidung nutzen.

Phase 5 Auszahlung

Aufgrund der abgegebenen verbindlichen Spielzüge wird die Auszahlung an die Agenten berechnet. Dabei wird die Ergebnismatrix des Gefangenendilemmas (siehe Abbildung 9) zugrundegelegt. Agenten die keinen Spielpartner gefunden haben, erhalten ihren Einsatz zurück. Hat ein Spieler keine Punkte mehr, kann er an keinem Spiel mehr teilnehmen.

Der komplette Ablauf aller Phasen ist noch einmal in einem Interaktionsdiagramm in Abbildung 15 zusammengefaßt. Dabei ist zu beachten, daß die Phase 2 so oft wiederholt wird, bis jeder Agent einen Spielpartner war oder maximal einmal Manager war.

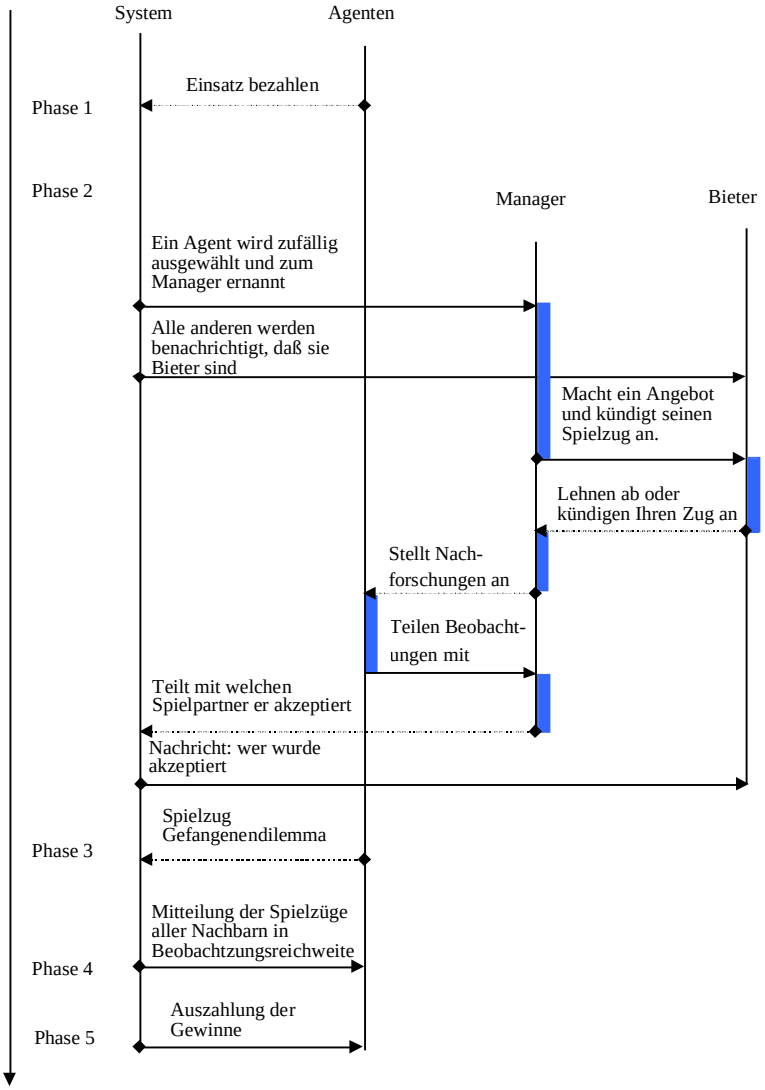

Abbildung 15: Vollständiges Protokoll für das Testbett

4.2.2. Wichtige Eigenschaften

In diesem Abschnitt wird das im vorhergehenden Abschnitt beschriebene Testbett charakterisiert. Wir zeigen außerdem, wie wichtige Eigenschaften der Anwendungsszenarien aus dem letzten Kapitel realisiert sind.

Spieltheorie. Dieses Testbett ist spieltheoretisch fundiert, damit Vergleiche zu Arbeiten auf diesem Gebiet möglich sind und Ergebnisse dieser Arbeit in vorhandene Ergebnisse eingefügt werden können. Insbesondere dient das *Gefangenendilemma* als Grundlage für Interaktionen. Damit lassen sich relativ einfach auch sehr komplexe Interaktionen, die zwischen Kooperation und Nichtkooperation unterscheiden, entsprechend den Szenarien modellieren. Ebenfalls wird damit der Gegensatz zwischen langfristigem Gewinn aus Kooperation und kurzfristigem Profit aus Ausbeutung erfaßt.

Ressourcen. Die Spieler besitzen ein beschränktes Maß an Punkten, die als abstrakte Ressource dienen. Sie befinden sich in einer Konkurrenz um weitere Ressourcen. Besonders schwierig ist dies, wenn, wie in den Anwendungsszenarien, nur wenige Interaktionen stattfinden, denn dann ist es sehr schwer zuverlässige Modelle aufzustellen. Da die modellierten Agenten sich probabilistisch bezüglich sozialer Rollen verhalten ist die Modellbildung weiter erschwert. Im Gegensatz zum klassischen Gefangenendilemma kostet schon die Teilnahme an einer Interaktion einen Teil dieser Ressource. Diese muß eingesetzt werden um mehr Ressourcen erwerben zu können und entspricht den Investitionskosten, die vor einem Vertragsabschluß stehen.

Multi-Agenten System. Für die Modellierung der Anwendungsszenarien sind Autonomie der Agenten, Skalierbarkeit, Kommunikation, Parallelität und Robustheit sehr wichtige Anforderungen. Daher ist die Experimentalumgebung als Multi-Agenten System realisiert.

Autonomie. Jeder der Agenten handelt autonom, in dem Sinne, daß er selbst Schlußfolgerungen aus seinen Beobachtungen zieht und diese benutzt, um sein Handeln zu planen. Dies soll auch ohne Interaktion mit dem Benutzer des Agenten möglich sein. Speziell für die beschriebenen Szenarien ist es wünschenswert, daß die Agenten auch selbst über die Wahl ihrer Interaktionspartner entscheiden können. Das Testbett ermöglicht daher, daß mehrere Agenten sich für eine Interaktion anbieten und der Agent sich für einen von ihnen entscheiden kann.

Lernen. Die beschriebenen Szenarien zeichnen sich durch die große Anzahl potentieller Akteure und die grundsätzlich mögliche Unbekanntheit ihrer Motive aus. Damit das Testbett auf die Szenarien paßt, muß es ermöglichen, daß keiner der Agenten von vorneherein weiß, welche Strategien einer der anderen spielen wird. Auch darf nicht bekannt sein, inwiefern die kommunizierten Daten der Wahrheit entsprechen. Jeder Spieler muß erst Ressourcen darauf verwenden, die anderen Spieler kennenzulernen, um Beobachtungen zu sammeln.

Kommunikation. Die Szenarien zeichnen sich dadurch aus, daß Kommunikation zwischen den Akteuren möglich und notwendig ist. Ein auszeichnendes Merkmal der beschriebenen Anwendungsszenarien ist es, daß den Agenten Wissen über ihre Mitspieler fehlt. Die Lösung dieses Problems mit Hilfe der Modellierung von

Vertrauen ermöglicht es, Kommunikation zu nutzen, um Beobachtungen anderer in Erfahrung zu bringen. Auch das Problem, daß der Mitteilende möglicherweise in seinen Mitteilungen lügt, und daß er selbst über nur eine möglicherweise nicht repräsentative Sammlung von Daten verfügt, wird berücksichtigt. Dies ist ein Verhalten, daß in den Szenarien auch von menschlichen Akteuren erwartet wird. Zum Beispiel würde man erwarten, daß ein Kunde vor einem Vertragsabschluß mit einer Firma sich zuerst über diese Firma informiert. Dies kann durch Befragung anderer Kunden geschehen. Die Modellierung umfaßt auch Kommunikation im weiterem Sinne, den Gebrauch von Presse und Internetdiensten. Ziel ist es, auch in Offenen Systemen Kommunikation zu nutzen, um die Verläßlichkeit von Daten zu bewerten.

Lügen, Betrügen. Das Testbett gewährleistet, daß jeder der Akteure andere über seine wahren Absichten täuschen kann und deren Unwissenheit zu seinem Vorteil nutzt. Bietet sich ein Akteur als Spielpartner an, ist es möglich, daß er Absichten verkündet, die er gar nicht einhalten will. Dies modelliert unter anderem, daß im *Electronic Commerce* Scheinfirmen gegründet werden, die dann Verträge abschließen und Leistungen entgegennehmen ohne Gegenleistungen zu erbringen.

Gerade bei der Benutzung vom Wissen anderer besteht das Problem, daß diese Intentionen haben, die möglicherweise dahingehen, andere über ihr Wissen zu täuschen. Beispielsweise wollen sie sich selbst als Interaktionspartner interessant machen. Dadurch haben sie kein Interesse daran, ihr Wissen über andere, eventuell bessere Spielpartner weiterzugeben. Desweiteren wollen sie als sehr glaubwürdiger Zeuge dastehen und werden versuchen, andere in Frage kommende Zeugen als unehrlich zu schildern.

4.2.3.Bezug zur praktischen Anwendung am Beispiel Virtueller Märkte

Nachdem wir in den letzen Abschnitten die Experimentalumgebung beschrieben und eine Übersicht über ihre Eigenschaften gegeben haben, soll nun an einem Beispiel ihr Bezug zu den Anwendungsszenarien detailliert dargelegt werden. Wir beschreiben an dieser Stelle den direkte Zusammenhang zwischen der Experimentalumgebung und einer Anwendung auf den *Electronic Commerce*.

Gegenüberstellung

In der linken Spalte stellen wir die praktische Anwendung der Virtuellen Märkten (*Electronic Commerce*) vor.	In der rechten Spalte stellen wir dar, in welcher Hinsicht das *Offen Gespielte Gefangenendilemma mit Partnerauswahl* dem entspricht.
Szenario	
Agenten sollen autonom für ihre Benutzer mit anderen Akteuren auf dem Virtuellen Marktplatz interagieren. Dabei kann es sich um den Erwerb von Waren oder Dienst-	Das vorgeschlagene System besteht aus autonomen Agenten, die zur Gewinnmaximierung mit anderen Agenten interagieren. Diese Agenten sind in der Lage mit dem Konzept des

leistungen handeln. Diese Interaktion ist unter Umständen auch ohne Zustandekommen des Austausches von Ware oder Geld mit Kosten verbunden.

Agenten

Es gibt Agenten, die unabhängig voneinander und zum Teil unkooperativ handeln. Jeder Agent will seinen Gewinn optimieren. Will ein Agent seinen Gewinn erhöhen, ist er möglicherweise auf die Kooperation mit anderen angewiesen. Diese Geschäftspartner sind weder notwendigerweise benevolent (altruistisch) noch zuverlässig.

Betätigungsfeld

Im internationalen Handel und insbesondere in virtuellen Marktplätzen gibt es nur wenige Strukturen, die das Einhalten von Vertragsversprechungen garantieren oder fördern. Es fehlen oft international legitimierte und effizient arbeitende Kontroll- und Strafmechanismen. Im allgemeinen wird das nicht-Einhalten eines Vertrages also nicht notwendigerweise geahndet.

Sozialer Druck

Vertragspartner, die häufig Verträge brechen, werden für andere unattraktiv, da sie sehr hohe Risiken bergen. Da die Anzahl der konkurrierenden Agenten sehr hoch ist und es in der Zusammensetzung des Marktes eine hohe Fluktuation gibt, ist es oft nicht möglich einen Vertragspartner einzuschätzen.

Selbst die Bewertung von Agenten durch Dritte ist mit Risiken behaftet: Diese Dritte handeln nach eigenen Motiven.

Vertrauens umzugehen. Sie benutzen es, um Informationen über andere Agenten zu sammeln und deren Unsicherheit in die Analyse einzubeziehen.

Dies wird identisch im Spiel abgebildet. Auch hier gibt es Abmachungen (Verträge) zwischen Agenten, die möglicherweise von einem der Vertragspartner unilateral gebrochen werden und zu einem erhöhten Verlust des anderen Agenten führen. Der Mechanismus, der dieses Problem in den Griff bekommt, ist die Berechnung des Vertrauens von Agenten in andere.

Auch hier gibt es keine direkte Strafe die ein Motiv für das Einhalten eines Vertrages darstellen würde. Die Experimentalumgebung ermöglicht es den Agenten über ihre Intentionen zu lügen. Durch die Kommunikation zwischen Agenten über das Verhalten von anderen kommt es jedoch zu schnellen Approximation des Verhaltens, auch ohne *a priori* Wissen. Dadurch werden egoistische und betrügerische Agenten schneller erkannt.

Es gibt keine *Trusted Third Parties*. Diese sind für manche Bereiche unrealistisch, da die Anwendung vielleicht die Kosten nicht rechtfertigt, bzw. die politische Grundlage auf der sie zu absolut vertrauenswürdigen Zeugen würden, fehlt.

Motive: Gewinnmaximierung

Alle Agenten sind daran interessiert, daß sie einen Gewinn aus den laufenden Geschäften beziehen. Dieser Gewinn kann daraus resultieren, daß sie ein Geschäft abschließen, welches für beide von Nutzen ist, oder aber zu Lasten ihres Vertragspartners geht.

Dieser Zwiespalt wird durch ein Spiel simuliert, das auf dem Gefangenendilemma beruht (Offen Gespieltes Gefangenendilemma).

Motive: Betrügen

Wird ein Agent von einem potentiellen Kunden über einen Konkurrenten befragt, so liegt es in seinem Interesse, nur die negativen Eigenschaften des Konkurrenten hervorzuheben.

Dies wird im System genau so wiedergegeben. Agenten können Informationen über andere Spieler austauschen.

Je nachdem wie ehrlich sie sind verschweigen sie jedoch Ereignisse, bei denen ihre Konkurrenten als zuverlässig oder altruistisch aufgefallen sind.

Resümee

Virtuelle Märkte sind gefährliche und unsichere Orte um Verträge abzuschließen, solange keine Strukturen für Kontrolle und Strafe vorhanden sind.

Die Berechnungen der Agenten erlauben es, das Verhalten anderer zu bestimmen. Durch diese Möglichkeit ergibt sich ein Mechanismus von Kontrolle und Strafe: Die Beobachtung von anderen Agenten, die Weitergabe dieser Beobachtung und die Auswahl der Spielpartner aufgrund gewichteter Beobachtungen. Spieler, die sich für andere Spieler gefährlich verhalten, werden dadurch bestraft, daß sie keine lukrativen Spielangebote mehr erhalten.

4.3. Exkurs: Sozionische Aspekte

Die Beobachtung und Analyse des Verhaltens der Mitspieler in einer oder mehreren Spielrunden stellt eine Aktivität in einem sozialen Kontext dar, deren Ergebnisse genutzt werden, um das zu Anfang rudimentäre Modell über andere Akteure in seiner Umwelt zu erweitern und zu verfeinern. Das Spielen des Offen Gespielten Gefangenendilemmas erlaubt sowohl Vortäuschung als auch Einhaltung sozialer Rollen. Wie schon das zugrundeliegende Gefangenendilemma ist dieses Modell also durchaus auch interessant für gesellschaftswissenschaftliche Forschung.

Deshalb werden wir in diesem Abschnitt einige soziologische Begriffe in Bezug zum vorgestellten Problem und seiner Lösung setzen. Dabei ist dieser Abschnitt weder vollständig noch kann er Anspruch darauf erheben, daß die soziologische Terminologie absolut korrekt wiedergegeben ist. Wir benutzen die Begriffe mit dem

Ziel, abstrakte Modelle verständlicher beschreiben zu können als dies mit reinen Informatikbegriffen möglich wäre. Dabei betrachten wir die von der Soziologie erfaßten Zusammenhänge, um neue Operationalisierungen für Multi-Agenten Systeme zu gewinnen. Daß diese Art der Metaphernnutzung in der Soziologie auf Widerstand stoßen kann, ist nicht verwunderlich, werden doch soziologische Fachtermini auf einen Bruchteil ihrer breiten Bedeutungspalette reduziert. Von der Verwendung dieser Metaphern erhoffen wir uns einen „fruchtbaren Innovationsbeitrag“ (siehe auch (Malsch et al., 1996)).

4.3.1. Vertrauen und Macht

In dieser Arbeit definieren wir den Begriff Vertrauen durch die beiden Konzepte Altruismus und Ehrlichkeit definiert. Ehrlichkeit ist zum einen definiert über das Betrügen im Informationsaustausch und zum anderen über das Betrügen im Spiel. Wir verabschieden uns bei diesem Ansatz also von der häufig anzutreffenden impliziten Annahme, daß Agenten benevolent bzw. kooperationswillig sind. Das Betrügen äußert sich in der Diskrepanz zwischen angekündigter Intention und ausgeführter Handlung. *Macht* spielt zwischen den Agenten eine wichtige Rolle. Der Zeuge besitzt die Macht über Information, die der andere Agent dringend benötigt. Für die Entscheidung, ob ein Agent als Spielpartner in Frage kommt, evaluiert der Entscheider sein in ihn gesetztes Vertrauen. In beiden Fällen dient Vertrauen dem Entscheider dazu, abzuwägen, wem er Macht über seine Ressourcen geben soll.

Vertrauen spielt in diesem Modell eine wichtige Rolle als Handlungskoordinationsmechanismus, wie er auch schon von Luhmann gesehen wurde (Luhmann, 1979). Vertrauen dient der Reduktion der Komplexität der Welt. Wer vertraut, geht aber auch gleichzeitig immer ein Risiko ein. Die Interpretation von Vertrauen geht in die Richtung von Luhmann und Deutsch: Vertrauen wird (im Vergleich zu anderen Definitionen) als „Risiko eingehen“ betrachtet. Dabei spielt weniger eine Rolle, sich auf etwas zu „verlassen“, als Wahrscheinlichkeiten und erwartete Gewinne und Verluste zu kalkulieren (Deutsch, 1973). Wie Bachmann feststellt, ermöglicht Vertrauen den Agenten spezifische Annahmen über das zukünftige Verhalten des Gegenübers zu machen und reduziert damit die Komplexität eines Handlungssystems (Bachmann, 1998). Dies trifft ganz offensichtlich für das Vertrauen in einen Zeugen zu: Jede Zeugenaussage hilft, das Modell über andere zu verbessern und damit deren Verhalten immer besser vorherzusagen.

4.3.2. Bestrafung durch Isolation

Hält sich ein Agent *A* grundsätzlich nicht oder nur selten an seine bekanntgegebenen Züge, und seine Mitspieler erleiden dadurch Nachteile, kann sich das für *A* negativ auswirken. Es kann passieren, daß er während der Partnerauswahl nicht mehr als Spielpartner ausgewählt wird; *A* wird durch Isolation bestraft. Da die Agenten zur Partnerwahl ihr Vertrauensmodell nutzen und dieses durch eigene Erfahrung sowie durch Kommunikation mit anderen während der Partnerauswahl erweitern und verfeinern (etwa im Sinne von „Ich habe hier ein Spielangebot von *A*, kenne ihn aber nicht, was kannst Du mir über ihn berichten?“), wirkt sich unkooperatives Spielen relativ schnell durch Isolation aus.

Wird der Agent von den anderen Mitgliedern der Gesellschaft sozial geschnitten, ist dies für ihn unter Umständen eine harte Bestrafung: Er kann an weiteren Spielen nicht mehr teilnehmen. Er wird ausgegrenzt und solange keine neuen Agenten in die Gesellschaft eintreten, verliert er die Möglichkeit sich wieder einzugliedern. Auch hier spielt Macht eine Rolle. Agenten erhalten Ressourcen nur dann, wenn ein anderer Agent ihnen mitzuspielen *erlaubt*.

4.3.3. Interaktion und Rollenverhalten

Die Agenten befinden sich in einer heterogenen Agentengesellschaft. Damit haben sie zwei einander möglicherweise widerstrebende Ziele. Einerseits ist jeder Agent daran interessiert, sein eigenes Überleben zu sichern und möglichst gut bei jedem Spiel abzuschneiden. Andererseits soll auch das altruistische Normverhalten eingehalten werden, um von der Gruppenzugehörigkeit zu profitieren und von vielen Agenten als Kooperationspartner erwünscht zu sein. Einem Agenten sind die Verhaltensmuster der anderen Agenten nicht bekannt, und um das Erkennen dieser Muster zu erschweren, verhalten sich die Agenten mit einer gewissen Wahrscheinlichkeit nicht rollenkonform (weiche Verhaltensmuster). Wird das soziale Verhaltensmuster des *Altruismus* konsequent angewendet, so erfahren Agenten viele Interaktionsangebote durch andere, da sie sich als gewinnbringende Interaktionspartner erwiesen haben. *Egoisten* schätzen das eigene Überleben und damit die eigene Punktzahl grundsätzlich wichtiger ein, als das Gesamtergebnis der Gesellschaft zu der sie gehören. Sie wählen im Gefangenendilemma immer den Spielzug, der am gewinnbringendsten scheint—Verrat; sie haben grundsätzlich kein Vertrauen in die Kooperationsbereitschaft anderer.

Kapitel 5
Vertrauen als ein berechenbares Konzept

„The problem is not trust… the problem is how he will *implement* what has been agreed upon."

Yasir Arafat über die Vertrauenswürdigkeit von Benjamin Netanyahu, 1997.

In diesem Kapitel beschreiben wir die Implementierung des im letzten Kapitel vorgestellten Formalismus. Im Zentrum dieser Beschreibung liegt die Implementierung einer Datenstruktur („`TrustNet`") im ersten Abschnitt. Diese Datenstruktur führt die Berechnungen durch, die zur Bestimmung des Vertrauens in Kooperationspartner und Zeugen notwendig sind. Dies schließt die Evaluation von beliebig vielen Aussagen über einen Agenten und den rekursiven Aufruf dieser Evaluation auf die Zeugen, die diese Aussagen gemacht haben, ein. Dabei werden die im letzten Kapitel vorgestellten Gleichungen angewendet. Wir zeigen, wie diese Datenstruktur algorithmisch bestimmt wird, wie sie mit Wahrscheinlichkeiten annotiert wird und wie sie Anfragen über die Vertrauenswürdigkeit eventueller Kooperationspartner beantwortet. Dieser Abschnitt stellt also die Lösung für das in Kapitel 3 beschriebene Problem dar. Ein Teil dieser Darstellung befindet sich auch in den Proceedings zum Workshop „Deception, Fraud and Trust" der *Autonomous Agents '99* Konferenz (Schillo et al., 1999b). Der zweite Abschnitt dieses Kapitels geht auf die Wahl des implementierten Algorithmus ein. Wir erläutern, welche anderen Ansätze erwogen und warum sie nicht verwendet wurden. Der dritte Abschnitt dokumentiert das Verhalten der Agenten in den verschiedenen Situationen, in denen sie im *Offen Gespielten Gefangenendilemma mit Partnerwahl* Entscheidungen treffen müssen. Den Abschluß des Kapitels bilden einige kurze Bemerkungen zur technischen Realisierung.

5.1. Das `TrustNet`: Berechnung von Vertrauen

Bei dem *Offen Gespielten Gefangenendilemma mit Partnerauswahl* sind für eine erfolgreiche Teilnahme an mehreren Stellen Entscheidungen notwendig. Diese Entscheidungen sind zum einen um so qualifizierter, je besser ein Agent sein Vertrauen in einen Agenten, der ihm Informationen übermittelt, berechnen kann. Zum anderen optimiert eine realistische Abschätzung des Altruismus eines potentiellen Spielpartners das eigene Spielergebnis eines Agenten A. Mit anderen Worten die Entscheidung von A wird qualifizierter, wenn er weiß wie hoch die Wahrscheinlichkeit ist, daß Q sich altruistisch verhält. Diese Wahrscheinlichkeit berechnet der Agent entweder selbst aus seinen eigenen Beobachtungen oder aber er fragt andere Agenten, wie sie Q einschätzen (dazu ist es wichtig die Vertrauenswürdigkeit des Informanten bestimmen zu können, s.o.). Letztendlich basieren die Einschätzungen aber immer auf einem empirischen Wert, der von einem beobachtenden Agenten aus den Verhaltensmustern von Q berechnet wird.

Es stehen dem Agenten aber nicht immer eigene Beobachtungen für eine solche Berechnung zur Verfügung. Daher soll er nach Möglichkeit mit anderen Agenten kommunizieren um von so vielen Erfahrungen wie möglich zu profitieren. Er muß aber damit rechnen, daß die Motive dieser Agenten derart sind, daß sie kein Interesse daran haben, ihre Erfahrungen ehrlich mit ihm zu teilen. Um das vorgestellte Vertrauensmodell für praktische Anwendungen interessant zu machen, muß der Agent also mit betrügenden Zeugen umgehen können. Es genügt nicht, daß er die Aussagen von eventuell betrügerischen Agenten ignoriert. Agenten, die absolut vertrauenswürdig sind, gibt es möglicherweise gar nicht.

Das Maß des Vertrauens eines Agenten in einen anderen hängt gemäß der im letzten Kapitel vorgestellten Definition von der Ehrlichkeit und des Altruismus eines Agenten ab. Im folgenden wird eine Datenstruktur vorgestellt, mit der zunächst die Ehrlichkeit eines Agenten berechnet werden kann. Der selbe Mechanismus kann angewendet werden, um seinen Altruismus zu bestimmen. Sind diese Werte bestimmt, kann nach Definition 17 auch die Vertrauenswürdigkeit berechnet werden.

5.1.1. Datenstruktur

In dem vorgeschlagenen Testbett fallen für jeden Agenten Entscheidungen an, bei denen er mit unscharfem Wissen, bedingten Wahrscheinlichkeiten und der Revidierung von bisherigem Wissen konfrontiert ist. Zur Berechnung dieser Entscheidungen wird eine einfache Datenstruktur, ein Graph, benutzt. Dieser eignet sich für die Lösung des beschriebenen Problems, da hiermit auf einfache Weise die Relationen zwischen den Zeugen und ihren Aussagen über andere ausgedrückt werden können. Außerdem läßt der Graph eine einfache Berechnung von Anfragen zu. Weiterhin bietet er nicht nur effiziente Techniken zum Einfügen neuer Daten in die bestehende Wissensbasis sondern auch eine anschauliche Modellierung. In Anlehnung an die Implementierung soll die besondere Klasse von Graphen, die hier besprochen wird, im weiteren als `TrustNet` bezeichnet werden.

Semantik

Im Wesentlichen handelt es sich bei einem Graphen um eine Menge von *Knoten* und eine Menge von *Kanten*. Im `TrustNet` repräsentieren die Knoten das *Modell* des Besitzers über einen Agenten. Die Kanten sind annotiert mit den *Beobachtungen* die von einem Zeugen über einen anderen Agenten mitgeteilt wurden. Dabei beziehen sich die Aussagen sowohl auf die Ehrlichkeit eines Agenten, als auch auf die Wahrscheinlichkeit eines der beschrieben sozialen Verhaltensmuster Altruismus und Egoismus. Für den Begriff *Beobachtung* wird Definition 11 angewendet.

Der Aufbau eines solchen Netzes ist an einem Beispiel in Abbildung 16 (links) demonstriert. Dieses Netz gibt den Wissensstand eines Agenten *X* wieder. Die Knoten repräsentieren sein Wissen über die Agenten *A*, *B*, *C* und *D*. Die Kanten drücken aus, daß von einem Agenten Aussagen über einen anderen gemacht wurden. Die Agenten *A* und *B* haben ihm Informationen über *C* zukommen lassen. *X* selbst hat Daten über *A* und *B* gesammelt, kann also selbst Aussagen über sie machen und damit auch die von ihnen geschickten Angaben einschätzen. Ein weiterer Agent *D* hat *X* Daten über *C* geschickt. Diesen Agenten kann *X* aber nicht aufgrund selbst gemachter Beobachtungen beurteilen. Die einzige Quelle für Daten über *D* ist *A*.

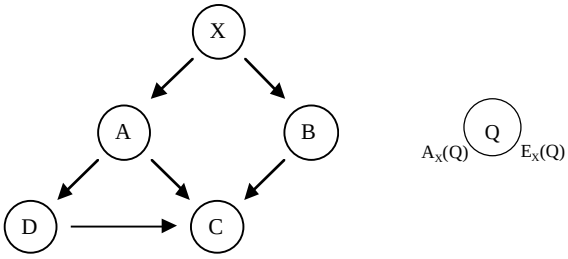

Abbildung 16: Ein einfaches `TrustNet` und Annotation der Knoten

In der Implementierung werden die mitgeteilten Beobachtungen mit den Kanten des Graphs gespeichert. Bei einer Auswertung werden sie benutzt, um den Altruismus und die Ehrlichkeit eines Agenten zu approximieren. Wie in Abbildung 16 (rechts) gezeigt werden die ermittelten Modelle (siehe Definition 15) den Knoten des `TrustNet` als Annotationen hinzugefügt. Die Datenstruktur liefert also einen einfachen Zugriff auf die Modelle.

Aufbau der Topologie des `TrustNet`

Jeder Agent hat eine eigene Datenstrukur, die seine „Sicht" der Welt beschreibt. Er beginnt mit einer Datenstruktur, einem Graphen, die nur aus einem Knoten besteht. Dieser Knoten repräsentiert ihn selbst. Wann immer der Agent einen anderen Agenten beobachtet, fügt er für diesen Agenten einen Knoten hinzu und annotiert die Kante mit dieser Beobachtung. Existiert der Knoten schon, wird dieser Knoten benutzt und kein neuer eingefügt. Existiert die Kante vom Agenten zum beobachteten Spieler schon, wird ihre Annotation um die gerade gemachte Beobachtung erweitert.

Gleiches gilt für Informationen, die den Agenten nicht durch Beobachtung, sondern durch Kommunikation von einem Zeugen erreichen. Auch sie werden als Annotationen von Kanten in den Graph eingefügt. Nur mit dem Unterschied, daß die Kante nicht vom Agenten selbst, sondern von seinem Informanten (dem

Zeugen) ausgeht. Da er nur Informationen bewerten kann, bei denen er auch den Informanten einschätzen kann, wird er also nur Agenten befragen, zu denen im `TrustNet` bereits ein Pfad existiert. Durch dieses Verfahren werden nur Knoten eingefügt, die auch eine eingehende Kante haben. Folglich gibt es in jedem `TrustNet` zu jedem Knoten einen Pfad von der Wurzel, die den Agenten selbst repräsentiert. Aufgrund der Möglichkeit mit beliebigen Agenten zu kommunizieren, ist es möglich, daß Zyklen im Graphen entstehen. Auf diese wird in Abschnitt 5.1.6 eingegangen. Wir können im weiteren davon ausgehen, daß die Agenten nur zyklenfreie Graph erzeugen.

5.1.2. Kombinieren von Aussagen mehrerer Zeugen

Nehmen wir an, daß ein Agent *Q* im `TrustNet` von Agent *X* durch einen Knoten repräsentiert wird, der nur eine eingehende Kante hat. Dann ist aus den vorhergegangenen Abschnitten bereits klar, wie *X* das Modell über die Ehrlichkeit von *Q* approximieren kann. Nach Konstruktion ist diese Kante mit einer Menge von Tripeln (siehe Definition 11) annotiert, die Aussagen über sein Verhalten in einer Menge von Spielen machen. Der Agent *X* muß also nur noch den Anteil von ehrlichen Spielen an diesen Spielen berechnen.

Besitzt der Entscheider Aussagen mehrerer Zeugen, ergibt sich für ihn das Problem, wie er diese Aussagen zusammenführt. Das Problem, warum einige Ansätze zu paradoxem Verhalten führen, wurde in Abschnitt 3.2 besprochen. An dieser Stelle sei noch einmal betont, daß es nicht genügt, die Häufigkeit eines Verhalten des Zielagenten für jede Zeugenaussage zu berechnen und diese Brüche dann zu mitteln. Es muß davon ausgegangen werden, daß sich die Beobachtungen teilweise überschneiden, eine Tatsache die bei einfachen arithmetischen Operationen nicht berücksichtigt würde. Im folgenden soll nun dargestellt werden, wie diese Informationen im Einklang mit der Wahrscheinlichkeitstheorie miteinander verrechnet werden können. Dabei wird der Fehler minimiert, indem Aussagen von mehreren Zeugen kombiniert werden. Dies ermöglicht gleichzeitig, daß der Entscheider mit der eventuell nur partiellen Information der Zeugen umgeht. Zur Vereinfachung betrachten wir zunächst nur die Berechnung der Ehrlichkeit eines Zielagenten. Danach verallgemeinern wir die Berechnung auf den Altruismus.

Betrachten wir nun zunächst den Fall, daß ein Entscheider seine eigenen Beobachtungen und die Aussagen von Zeugen betrachtet, deren Ehrlichkeitswerte er schon bestimmt hat. Der Entscheider hat demnach den Parameter *p* der Funktion *Betrügen* für jeden dieser Agenten (mehr oder weniger gut) approximiert. Aber wie soll er nun die Beobachtungen von mehreren Zeugen (und seine eigenen) miteinander kombinieren?

Hat der Entscheider anhand seines Ehlichkeitsmodells der Zeugen approximiert, welche Zeugen über wieviel Beobachtungen gelogen haben, dann kann er wie folgt eine Tabelle aufstellen. In dieser Tabelle ist für jeden Zeugen dessen Aussage und eine Schätzung von dem, was er verheimlicht hat, aufgetragen. Ein Beispiel einer solchen Tabelle findet sich in Abbildung 17. Ein „√“ bedeutet dabei eine Beobachtung eines positiven, ein „*X*“ eine Beobachtung eines negativen Verhaltens. Ein Fehlen eines Symbols steht für das Fehlen (oder Verheimlichen) einer

Beobachtung für diese Spielrunde. Anhand der mitgeteilten Beobachtungen kann der Entscheider ein Ergebnistupel zusammensetzen, daß für jede Spielrunde nur ein Ergebnis enthält. Somit ist das Problem der mehrfachen Gewichtung von Beobachtungen gelöst. Es bleibt noch die Frage, wie der Agent das Wissen über den Umfang des Betrugs durch die Zeugen einsetzt, um das tatsächlich stattgefundene Verhalten zu approximieren.

Spielrunde	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Betrogen
Zeuge 1				√		√			X		X	X		X	**0,7**
Zeuge 2		√		√			X		X	X					**1,1**
Ergebnistupel		√		√		√	X		X	X	X	X		X	**?**

Abbildung 17: Zusammensetzung zweier Zeugenaussagen

Es bieten sich mehrere Möglichkeiten: Die Anzahl der verschwiegenen „*e*"s könnte gemittelt oder addiert werden, es könnte das Maximum oder das Minimum gewählt werden. Diese Ansätze entscheiden sich aber sehr willkürlich für eine Berechnung. Im folgenden soll eine Heuristik angegeben werden, die sich auf eine stochastische Begründung stützt.

Die grundsätzliche Schwierigkeit besteht darin, die mögliche Überdeckung der nicht mitgeteilten Beobachtungen zu approximieren. Betrachten wir die Tupel, die dem Entscheider von den Zeugen mitgeteilt wurden, so läßt sich die Überdeckung dieser Einträge leicht bestimmen. Da einzelne Einträge von betrügenden Zeugen zufällig weggelassen werden, kann angenommen werden, daß die Überdeckung der mitgeteilten Tupeleinträge *gleich* der Überdeckung der nicht mitgeteilten Tupeleinträge ist. Das Maß dieser Überdeckung wird in folgender Definition als *Dichte* definiert.

Definition 19: Dichte von Tupeleinträgen

Sei *Matrixeinträge* die Anzahl aller Einträge, die in allen Tupeln enthalten sind (man kann die Tabelle, die die Tupel aller Zeugen enthält, auch als Matrix betrachten). Seien weiterhin *Zeugen* die Anzahl der Tupel und *Tupellänge* die Anzahl der Einträge im Ergebnistupel (wie z.B. in Abbildung 17). Die *Dichte* der Einträge mehrerer Tupel ist dann definiert als:

$$Dichte = \frac{Matrixeinträge \,/\, Zeugen}{Tupellänge}. \qquad (16)$$

Sie gibt an, wie oft Zeugen über gleiche Spielrunden berichtet haben.

Unter obiger Annahme, daß die Dichte der übermittelten Einträge gleich der Dichte der nicht übermittelten Einträge ist, gilt dann:

$$Dichte = \frac{verschwiegene\ Einträge \,/\, Zeugen}{zusätzliche\ Tupellänge}. \qquad (17)$$

Es ist bekannt, daß für die verschwiegenen Einträge „*e*"s eingesetzt werden können (vgl. Abschnitt 4.1.3 zu den Motiven der Agenten). Wieviel das im Ergebnistupel sein müssen, gibt die Variable *zusätzliche Tupellänge* an. Wenn wir die beiden Formeln (16) und (17) gleichsetzen, ergibt sich für die zusätzliche Tupellänge folgender Wert:

$$zusätzliche\ Tupellänge = \frac{verschwiegene\ Einträge \,/\, Zeugen}{Dichte}. \qquad (18)$$

Diese Vorgehensweise (zusammengefaßt dargestellt in Abbildung 18) approximiert wesentlich besser das Modell aufgrund mehrerer Zeugenaussagen. Sie berücksichtigt, daß Einträge mehrfach auftreten können und löst damit das Problem aus Abschnitt 3.2.

Weiterhin ergibt die Untersuchung der Dichte von Einträgen eine bessere Approximation der Gesamtzahl der zu ergänzenden Information, da sie keine Annahmen über deren Verteilung über die Spielrunden macht, sondern eine qualifizierte Schätzung berechnet. Der Agent kann aufgrund des ermittelten Ergebnistupels und der zusätzlichen Tupellänge einen zuverlässigeren Anteil von ehrlichem Verhalten bestimmen. Damit ist die in Abschnitt 4.1.3 beschriebene Modellierung von Vertrauen in Zeugen vollständig berechenbar. Die Berechnung der zweiten Anwendung von Vertrauen, nämlich das Vertrauen in ein Kooperationsangebot von einem Interaktionspartner, behandeln wir im übernächsten Abschnitt. Zunächst soll noch auf die Approximation des Altruismus eines Agenten eingegangen werden.

5.1.3. Approximation von Verhalten am Beispiel Altruismus

Die Berechnung des Altruismus eines Agenten erfolgt analog zu obigem Verfahren. Wir gehen wieder davon aus, daß verschiedene Zeugen ihre Beobachtungen mitgeteilt haben. Nur betrachten wir jetzt nicht die Aussagen über Ehrlichkeit, die in diesen Beobachtungen steckt, sondern die Aussagen über den Altruismus eines Agenten. Dies bedeutet, daß wir diesmal nicht die Menge der übermittelten *e* für die Einschätzung des Zielagenten betrachtet, sondern die Aussagen über den Altruismus, also die Anzahl der übermittelten „*a*"s. Trotzdem muß die Ehrlichkeit der Zeugen vorher berechnet werden, damit eingeschätzt werden kann, wieviel Beobachtungen mit einem *a* sie wohl verschwiegen haben. Die Berechnung des Altruismus ist also direkt analog zu den Ehrlichkeitsberechnungen.

Im übrigen hat sich bei der Entwicklung dieses Formalismus herausgestellt, daß sich Zeugenaussagen über jedes beliebige Verhaltensmuster auf diese Weise behandeln lassen. Es können also Informationen beispielsweise über die Intelligenz eines Agenten, seine Geschwindigkeit, seine Verfügungsmöglichkeit über Ressourcen etc. kommuniziert werden. Voraussetzung dafür ist, daß die Information von Agenten beobachtet werden kann und daß sie sich durch eine Dichotomie (Begriffspaar) beschreiben läßt. Dies könnten zum Beispiel auch aussagen sein wie verfügt-

über-Ressource-X / verfügt-nicht-über-Ressource-X oder kann-A-erledigen / kann-A-nicht-erledigen. Schließlich muß eingeschätzt werden können, welche Informationen ein Zeuge durch die Betrügen-Funktion eliminieren würde. Damit ergibt sich, daß der vorgestellte Mechanismus wesentlich allgemeiner ist als beschrieben und vielseitig zur Kommunikation und Evaluation von Beobachtungen eingesetzt werden kann.

---

procedure *bestimmeEhrlichkeitVonAgent (q)*
begin
verschwiegeneEinträge = 0
foreach *Elternknoten z von q* **do**
bestimmeEhrlichkeitVonAgent (z)
foreach *Elternknoten z von q* **do**
begin

$$N = \left\{ r \,\middle|\, \begin{array}{l} r \text{ ist eine Spielrunde, in der q von z bei} \\ \text{ehrlichem Verhalten beobachtet wurde} \end{array} \right\}$$

if $\left(|N| = 0 \right)$

$$verschwiegeneEinträge \mathrel{+}= \frac{|N|}{ehrlichkeit(z)} - ehrlichkeit(z)$$

else

$$verschwiegeneEinträge \mathrel{+}= \frac{ln\ 1/2}{ln(1 - ehrlichkeit(z))}$$

end
dichte = Dichte des Ergebnistupels

$$zusätzlicheTupellänge = \frac{verschwiegeneEinträge}{dichte \cdot \left| \{ z \,|\, z \text{ Elternknoten von } q \} \right|}$$

$$E = \left\{ r \,\middle|\, \begin{array}{l} r \text{ ist eine Spielrunde, in der q bei} \\ \text{ehrlichem Verhalten beobachtet wurde} \end{array} \right\}$$

$$ehrlichkeit(q) = \frac{|E| + zusätzlicheTupellänge}{längeErgebnistupel + zusätzlicheTupellänge}$$

end

Abbildung 18: Algorithmus für die Berechnung der Ehrlichkeit eines Agenten unter Berücksichtigung der Vertrauenswürdigkeit der Zeugen

5.1.4. Vertrauen in Kooperationspartner

Nach Definition 16 ist die Einschätzung der Vertrauenswürdigkeit eines Agenten zusammengesetzt aus der Einschätzung zweier Verhaltensweisen: Ehrlichkeit und seinen Altruismus. Dabei wird die Vertrauenswürdigkeit eines Kooperationsversprechens betrachtet. Die Alternative eines Kooperationsversprechens ist die

Ankündigung von Verrat. In diesem Fall erübrigt sich die Überlegung, ob mit ihm kooperiert werden soll, schon durch die Analyse der Ergebnismatrix.

Hat ein Agent X ein Kooperationsversprechen von einem Agenten Q, dann benötigt er für eine qualifizierte Entscheidung über dieses Versprechen erst ein Modell des Verhaltens von Q. Legen wir Definition 15 zugrunde, so besteht dieses Modell in der Berechnung der Ehrlichkeit und des Altruismus. Wie diese Berechnungen erfolgen ist in den Abschnitten 5.1.2 und 0 angegeben. Sobald die Werte für das Modell gegeben sind, erfolgt die Berechnung für die Vertrauenswürdigkeit V von Q aus der Sicht von X direkt nach Definition 17 und durch Formel (15):

$$V_{X(Q)} = \frac{A_X(Q)}{A_X(Q) + (1 - A_X(Q))(1 - E_X(Q))}.$$

Ob X jetzt mit Q kooperiert, hängt davon ab, in welchem Maß X bereit ist ein Risiko einzugehen. Dies hängt aber nicht nur vom Agenten ab, sondern auch von dem was auf dem Spiel steht, also auch von dem Szenario, in dem gespielt wird. Auf die Risikoschwelle, die die implementierten Agenten benutzen, wird in Abschnitt 5.3.2 eingegangen.

5.1.5. Zeugenauswahl und Belief Revision

Ist ein Agent in der Phase, in der er andere nach ihren Beobachtungen fragen kann, so läßt er vom `TrustNet` die *n* ehrlichsten Agenten bestimmen. Die Antworten dieser Zeugen werden genauso wie die eigenen Beobachtungen in das `TrustNet` eingefügt. Bevor der Agent das Modell eines anderen Agenten aus seiner Datenstruktur ausliest, werden die notwendigen Neuberechnungen durchgeführt (siehe nächster Abschnitt). Mit dieser Neuberechnung wird also auch der *belief*, seine Einschätzung des anderen Agenten, neu evaluiert.

5.1.6. Zyklen im `TrustNet`

Zyklen bergen für diese Datenstruktur das Problem, daß sich eventuell selbstverstärkende Schleifen ergeben, die einen inferierten Wert unangemessen in die Höhe treiben oder unangemessen absenken. Beth et al. (1994) argumentieren, das es sicher ist, Kanten auszulassen, die solche Zyklen produzieren, da die Zyklen keine wesentliche Information enthalten. Trotzdem muß aber sorgfältig ausgewählt werden, was mit einer einzufügenden Kante getan werden muß, die einen Zyklus erzeugt. Würde die Kante ohne nähere Betrachtung ausgelassen, so könnte es sein, daß diese neue Kante mehr Information enthält als eine bereits im Zyklus existierende und man auf diesem Weg hilfreiche Information eliminiert. Unter Berücksichtigung der Tatsache, daß dieser Ansatz auf zeitlich veränderliches Verhalten erweiterbar sein soll, ist es z.B. auch wichtig nach einer gewissen Zeit ältere Kanten gegen neuere auszutauschen, um das Modell möglichst aktuell zu halten.

Um dieses Problem zu lösen wird nach der in Abbildung 19 dargestellten Methode vorgegangen. Macht ein Agent neue Beobachtungen, oder wird ihm eine Menge von Beobachtungen mitgeteilt, so entspricht dies einer in das `TrustNet` einzufügenden Kante. Jede solche Kante wird zunächst vorläufig in den Graph

eingefügt. Dann wird getestet, ob dadurch im Graph ein Zyklus entstanden ist. Ist dies nicht der Fall kann sie endgültig eingefügt werden. Für den Fall, daß ein Zyklus entstanden ist, wird getestet, ob dieser Zyklus eine Kante enthält, deren Informationsgehalt geringer ist, als der der eingefügten Kante. Dabei ist das Kriterium die *Menge* der Information. Dieses Kriterium kann aber auch in *Aktualität* der Information geändert werden. Gibt es eine solche Kante, wird überprüft, ob der Graph noch Zyklen enthält, wenn diese weniger informative Kante entfernt wird. Ist dies der Fall, wird die weniger informative Kante entfernt, die andere Kante wird endgültig eingefügt. Für den Fall, daß immer noch Zyklen enthalten sind, muß die vorläufig eingefügte Kante wieder entfernt werden. Die gefundene, weniger informative Kante wird trotzdem entfernt. Dies gewährleistet, daß mit der Zeit aussagekräftigere Kanten andere ersetzen und so das gesamte Netzwerk an Wissen dazugewinnt.

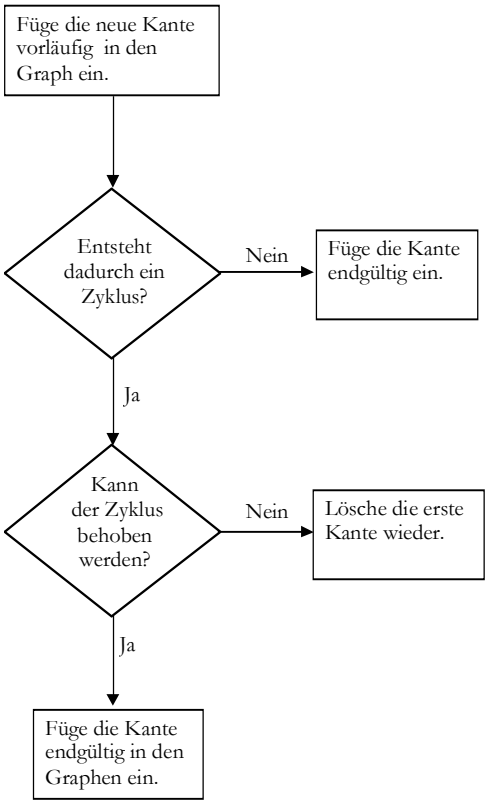

Abbildung 19: Erhaltung der Graphkonsistenz

5.1.7. Komplexität der Berechnungen des `TrustNet`

Die Komplexität der Berechnungen im `TrustNet` stellt sich folgendermaßen dar: Jeder Agent besitzt eine `TrustNet`-Datenstruktur. Diese enthält maximal so viele Knoten, wie es Agenten gibt, beschrieben durch die Variable *n*. Da die Agenten so viele Fragen an andere Agenten stellen wie nur möglich ist, sind die Graphen sehr dicht, selbst wenn sie keine Zyklen enthalten. Im schlimmsten Fall ist jeder Knoten mit fast jedem anderen Knoten durch eine Kante verbunden. Dann müssen wir von $O(n^2)$ Kanten pro Graph ausgehen. Die Kanten selbst sind mit einer Menge von

Beobachtungen annotiert. Diese Mengen von Beobachtungen wachsen theoretisch mit der Anzahl der gespielten Runden *r*. In dieser Arbeit haben wir darauf verzichtet, die Größe der gespeicherten Beobachtungsmenge zu beschränken, da hier insbesondere das Verhalten des `TrustNet`s mit sehr wenig Information betrachtet werden sollte. Würde der Ansatz für *beschränkt rationale* Agenten erweitert, so besteht die Möglichkeit, diese durch eine Einschränkung der Größe der Beobachtungsmengen auszudrücken. In Abhängigkeit der Anzahl der Agenten *n*, die in der Datenstruktur repräsentiert werden und den Runden *r*, über die Informationen vorliegen, ergibt sich also für die Komplexität der Berechnung aller Werte aller Knoten $O(n^2 r)$.

Um die Komplexität zu drücken, wird in der Implementierung des `TrustNet` eine Berechnung nicht bei Einfügen einer Kante ausgeführt (was relativ häufig vorkommt), sondern nur dann, wenn eine Anfrage über einen Agent beantwortet werden soll (*lazy evaluation*). Dann werden nur die Knoten neu berechnet, deren eingehende Kanten sich geändert haben, oder deren Vorgängerknoten neu berechnet werden müssen. Sind die Werte für einen Knoten neu berechnet, so werden sie am Knoten gespeichert. Dieser Knoten wird als berechnet markiert. Dies bedeutet, daß das ganze Netz nur im schlimmsten Fall ganz neu berechnet wird. Dieser schlimmste Fall tritt dann ein, wenn alle ausgehenden Kanten des Wurzelknotens geändert wurden und die geänderten Knoten alle anderen Knoten als Vorfahren im Graph haben.

Das Verfahren der *lazy evaluation* macht es im Gegenzug erforderlich, daß nach dem Einfügen einer Kante, der Unterbaum als neu zu berechnen markiert werden muß. Dieser Markierungsdurchlauf hat die Komplexität $O(n)$. Für das Einfügen einer Kante selbst wird ein Zyklentest, wie in Abschnitt 5.1.6 beschrieben, durchgeführt. Dieser hat die Komplexität $O(n^2)$. Danach wird die Kante als Zeiger bei beiden Knoten eingetragen, was fast konstante Zeit kostet und damit unter der Komplexität des Zyklentests liegt. Wie bereits gesagt, wird ein Markierungsdurchlauf notwendig, um die Komplexität des Neuberechnens zu verringern. Da die besuchten Knoten nicht identisch sind (falls Zyklen auftreten, werden beim Zyklentest auch Vorfahren eines Knotens besucht) ergibt dies $O(n^3)$ für das Einfügen von Kanten. Da phasenweise eingefügt und erst dann aus der Datenstrukur ausgelesen wird, würde ein Neuberechnen der Knoten während des Einfügens der Kanten zwar theoretisch zur selben Komplexität führen, praktisch aber die Effizienz verschlechtern.

Für den Zugriff auf die Werte der Knoten ohne Neuberechnung der Datenstruktur und das Einfügen eines Knotens wurden die in der Java™ Klassen-Bibliothek zur Verfügung gestellten Hash-Funktionen verwendet. Deren Komplexität ist in der Dokumentation nicht angegeben, aber es kann davon ausgegangen werden, daß sie im Rahmen üblicher Verfahren, also bei etwa $O(n \log n)$ liegt.

Zusammengefaßt heißt dies, daß die Komplexität

- für das Auslesen eines Wertes bei berechnetem `TrustNet` $O(n \log n)$,
- für das Einfügen eines Knotens $O(n \log n)$,
- für das Einfügen einer Kante $O(n^3)$ und
- für das Neuberechnen des ganzen Netzes $O(n^2 r)$ ist.

5.2. Rechtfertigung der Methodenwahl

Es wird darauf eingegangen, wie das Problem der Transitivität gelöst wurde, warum keine Bayes'schen Netze benutzt wurden und warum statt der verwendeten Graphentechnik andere prominente Techniken der Informatik in diesem Fall nicht benutzt wurden.

5.2.1. Transitivität von Vertrauen

Wie schon in Abschnitt 3.2 beschrieben, gibt es Literatur, die die Transitivität von Vertrauen anfechtet. Dieser Kritik wird in diesem Ansatz Rechnung getragen. Zunächst einmal folgt nicht aus „*A* vertraut *B*" und „*B* vertraut *C*", daß dann automatisch „*A* vertraut *C*" gilt. Der hier beschriebene Algorithmus berücksichtigt jeweils das Maß der subjektiven Vertrauenswürdigkeiten und die Motive aller Zeugen. Es werden die vorhandenen Daten nicht auf eine zweiwertige Logik reduziert, sondern es wird sowohl eine qualitative als auch eine quantitative Aussage über das Vertrauen von *A* in *C* gemacht.

5.2.2. Warum keine Bayes'schen Netze?

Es gibt eine Reihe von Arbeiten, die sich mit verteilten probalisitischem Schließen auseinandersetzen. Einem Verfahren also, bei dem es darum geht, aus einer Menge von Aussagen von verschiedenen Agenten, eine neue, qualifiziertere Aussage zu erhalten (z.B. (Biswas et al., 1999a), (Zeng und Sycara, 1996), (Xiang, 1994), (Schum, 1981)). Dies wird gerne mit Bayes'schen Netzen modelliert, weshalb die Frage naheliegt, warum dies nicht auch in der vorliegenden Arbeit getan wurde.

Für die Modellierung mit Bayes'schen Netzen spricht deren ökonomische Darstellung von abhängigen Wahrscheinlichkeiten, ihre einfache Modellierungssprache und elegante Spezifizierung der Funktion, die Informationen mehrerer Elternknoten kombiniert. Ein Bayes'sches Netz ist ein gerichteter Graph mit einer Menge von Knoten und einer Menge von Kanten. Dabei repräsentieren die Knoten Ereignisse und die Kanten kausale Abhängigkeiten. Zwei Kanten, die von verschiedenen Knoten ausgehen repräsentieren also voneinander unabhängige Einflüsse auf ein drittes Ereignis.

Darin besteht das für die beschriebenen Szenarien wichtigste Problem. Die Aussage eines Agenten *A* über einen anderen Agenten *Q* ist nicht notwendigerweise unabhängig von der Aussage eines Agenten *B* über *Q*. Die Korrelation zwischen beiden ist sehr schwierig, oder je nach Konstellation der mitgeteilten und nichtmitgeteilten Beobachtungen, gar nicht zu berechnen. Kausale Unabhängigkeiten und unabhängige Wahrscheinlichkeiten zwischen den Ereignissen sind aber für die Modellierung mit Bayes'schen Netzen zwingend notwendig ((Charniak, 1991), (Russell und Norvig, 1996), (Pearl, 1997). Diese Argumentation führte auch zur Ablehnung der Dempster-Shafer Theorie. Hinzu kommt, daß es schwer ist, zeitliche Veränderungen mit Bayes'schen Netzen angemessen zu erfassen. Bayes'sche Netze registrieren nur sehr schwer Veränderungen über die Zeit, da sie jedes neue Datum, welches in das Netz eingefügt wird, als gleichwertig betrachten. Dadurch entsteht ein Mittelwert, der neuere Daten in ihrer Bedeutung genauso wie ältere Daten behandelt

(Pearl, 1988). Dieses Problem ist z.B. in der Benutzermodellierung von großer Bedeutung (Preece, 1994).

Aus diesem Grund, und da der vorliegende Mechanismus für ein über die Zeit veränderliches Verhalten offen sein sollte, wurde diese Möglichkeit der Implementierung ausgeschlossen. Der Ansatz wurde allgemeiner gefaßt und die Kombinationsfunktion für die Informationen der Elternknoten selbst definiert. Dabei konnte die Eleganz der Darstellung von Abhängigkeiten der Daten (nicht notwendigerweise kausalen) mithilfe eines Graphen beibehalten werden. Daß die oben beschriebenen Ansätze mit Bayes'schen Netzen auskommen, liegt daran, daß sie zwar von Agenten mit unscharfem Wissen ausgehen, aber nicht in Betracht ziehen, daß diese Agenten bei der Kommunikation ihrer Informationen betrügen (Xiang, Schum und Biswas et al.) oder aber keine Kommunikation erlauben (Zeng und Sycara). Die Anwendung von Xiang z.B. ist die Kombination der Diagnosen dreier Ärzte. Zwar macht es in diesem Beispiel Sinn, deren Aussagen zusammenzufassen, in einem Offenen System kann jedoch nicht von benevolentem Verhalten ausgegangen werden.

5.2.3. Andere Berechnungsmethoden

Sicherlich hätte der in dieser Arbeit vorgeschlagene Algorithmus auch mithilfe anderer KI-Techniken als der Wahrscheinlichkeitsrechnung und einem Graphenalgorithmus realisiert werden können. Trotzdem fanden wir Argumente gegen einen anderen Ansatz und wir wollen diese im folgenden Abschnitt kurz zusammenfassen.

Nach Zimmermann (1995) gilt zumindest für das besprochene public-key Szenario, daß für das Maß des Vertrauens alle Werte von null bis eins möglich und notwendig sind. Gambetta argumentiert, genauso wie Marsh, daß Vetrauen notwendigerweise dieses breite Spektrum von Werten annehmen muß ((Gambetta, 1990b), (Marsh, 1994)). Damit scheiden die *Prädikatenlogik* und die dafür gängigen Kalküle als Modellierungssprache und Inferenzmechanismen aus.

Als Alternative kommt die *Modallogik* in Frage, da ihre Operatoren mit Bedeutungen belegt sind, die neben den Zuständen wahr und falsch auch Eventualität und Unmöglichkeit ausdrücken. In der Modallogik würde aber das Wissen über die Vertrauenswürdigkeit, das in sehr großer Detailliertheit vorliegt, auf den binären Zustand eines oder mehrerer zu definierender Modaloperatoren reduziert. Die Anzahl der Operatoren (und damit die Komplexität der Berechnung) würde mit dem Detaillgrad der Modellierung wachsen. Außerdem ist diese Information nur schwer in einem Kalkül zu benutzen. Wie bereits in Abschnitt 3.2 argumentiert, entzieht sich die Relation „Vertrauen" grundlegender Eigenschaften wie z.B. der Transitivität. Um adäquate Ergebnisse zu erhalten, sind außerdem komplexe numerische Berechnungen über graphenähnlichen Datenstrukturen notwendig.

Damit scheiden auch KL-ONE und ähnliche Repräsentationen und Inferenzsysteme wie CLASSIC und LOOM aus, die im wesentlichen zur Repräsentation von Wissen entworfen wurden, das eine terminologische Struktur hat. Die Betrachtung von Semantischen Netze wie z.B. Konzeptgraphen oder Frame-Systeme wie OWL, KODIAK zeigt, daß hier wie dort Modelle mit Hilfe von Graphen mit Knoten und

Kanten konstruiert werden. Dort sind Kanten aber definiert als Relationen zwischen den Knoten, die nach einer taxonomischen Struktur definiert sind. Ähnlich wie KL-ONE scheiden auch diese Systeme aus, da im Fall dieser Arbeit keine taxonomische Struktur vorliegt, sondern ein Geflecht aus Objekten gleicher Bedeutung und Relationen *verschieden starker Ausprägung*. Dies ist vergleichbar mit der Tatsache, daß hier zwar ein *belief* im klassischen Sinne berechnet wird („Ist Agent *X* vertrauenswürdig oder nicht?"), es während der Berechnung wichtig ist einzubeziehen, mit welchem Maß dieses Vertrauen vorhanden ist.

Von der Problembeschreibung her läßt sich also folgern, daß eine *belief revision* notwendiger Bestandteil einer Lösung sein muß und es sich hier um eine Anwendung von Probabilistischem Schließen handelt. Dies beinhaltet die Tatsache, daß eine zugrundeliegende Logik nicht-monotone Eigenschaften haben muß. Der Einwand, daß für das gegebene Problem ein probabilistischer Ansatz erforderlich ist, läßt auch *Defaultreasoning*, eine klassische nicht-monotone Logik, ausscheiden. Zwar bietet Defaultreasoning die Möglichkeit, mit Unwissen über andere Agenten umzugehen, voreingestellte Annahmen über unbekannte Agenten zu machen und diese nach Sammeln von Erfahrungen wieder zurückzunehmen. Es löst das Problem der Reduktion der gesammelten Daten auf ein binäres Prädikat aber nicht und kann auch nicht *a priori* mit dem komplexen Zusammenhang von Zeugenaussagen und den enthaltenen Betrügen umgehen.

Ein Versuch, das hier gelöste Problem mit *Künstlichen Neuronale Netzen* (*KNN*) zu lösen, erschien wenig erfolgversprechend. Zwar sind *KNN* in der Regel sehr gut darin, Approximationsfunktionen zu probabilistischem Verhalten zu lernen. Jedoch benötigen sie sehr viele Daten um komplexe Zusammenhänge korrekt zu approximieren (Freeman und Skapura, 1991). Wie in Abschnitt 3.3 beschrieben, steht aber nur eine geringe Datenmenge zur Verfügung. Begriffsverbände dagegen sind zu statisch, um sich den zur Verfügung stehenden Daten anzupassen (Wille, 1993). Obwohl den Agenten zum Wissenserwerb Attributvektoren zur Verfügung stehen, handelt es sich hier nicht um eine Form des fallbasierten Schließens. Während im Fallbasierten Schließen Attributvektoren klassifiziert werden, sollen in den beschriebenen Szenarien Vorhersagen über das zukünftige Verhalten des Produzenten solcher Vektoren eingeschätzt werden. Die Verhaltensweisen sind aber per Definition und durch die Ergebnismatrix schon klassifiziert. Es genügt also die gesammelten Fälle von tatsächlich erfolgtem Verhalten zu sammeln und zu bewerten.

Es ist jedoch vorstellbar, daß dieser Ansatz als Schnittstelle zu Inferenzsystemen der oben beschriebenen Art dient. So ist es möglich, einen Modaloperator „vertraut" zu definieren, dessen Wahrheitswert immer aus dem hier vorgestellten System ausgelesen und mittels eines Schwellenwertes in einen Boole'schen Wert umgewandelt werden muß. Zwar könnte dieses System dann nicht inferieren, welchem Agenten es wie sehr vertraut. Es wäre aber möglich diese Information für andere Inferenzen zu nutzen und in ein größeres System zu integrieren. Dadurch würden in diesem System weitere Wege zum Informationsgewinn zur Verfügung stehen. Ähnliches gilt für eine KL-ONE-artige Repräsentation.

In gewisser Weise wird der Ansatz der *Fuzzy Logic* benutzt, da der Begriff des „vertrauenswürdig sein" über ein von Agent zu Agent verschiedenes Intervall definiert ist und das Maß des Vertrauens dem Grad des Zutreffens dieses Attributs

entsprechen könnte. Es wurde wegen der Klarheit des in Abschnitt 0 beschriebenen Modells vorgezogen, die Datenstruktur und die Inferenz selbst zu implementieren. Als System stand dafür die am DFKI entwickelte Simulationsumgebung *Social Interaction Framework* (*SIF*) zur Verfügung (siehe auch Abschnitt 5.4).

5.2.4. Warum ist dieses Modell neu?

Wie schon in Kapitel 2 gezeigt, existieren wenige implementierungsfähige Ansätze zur Modellierung von Vertrauen. Zwei der anspruchsvolleren Ansätze wurden in Abschnitt 2.3.2 beschrieben. Diese behandeln jedoch beide nicht die Kommunikation mit anderen zu Beschleunigung und Verbesserung der Modellbildung. Um dies leisten zu können, benötigen sie eine weitere Verfeinerung. In den Tests von Marsh (1994) können die Agenten sich nicht beide ihren Interaktionspartner aussuchen. Ein Agent zwingt einem anderen eine Interaktion auf. Dadurch fehlt eine wichtige Voraussetzung für die Anwendbarkeit z.B. auf das *Electronic Commerce* Szenario, bei dem ja beide Akteure einer Interaktion (beispielsweise einem Kauf) zustimmen. Außerdem gibt es in seiner Analyse keinerlei Kommunikation zwischen den Agenten, in der sie ihre Intentionen ankündigen können. Obwohl er Vertrauen modelliert, bleibt so die Möglichkeit des Betruges unberücksichtigt.

Die Arbeiten von Castelfranchi et al. (1998) haben einen anderen Schwerpunkt. Sie gehen davon aus, daß eine Abschätzung der Wahrscheinlichkeit, daß die Intention eines anderen Agenten ehrlich gemeint ist, vorliegt. In dieser Arbeit wird dagegen gezeigt, wie a) diese Wahrscheinlichkeit durch Beobachtung angenähert werden kann und wie b) die Kommunikation mit (möglicherweise) betrügerischen Agenten dazu dient, dieses Modell schneller zu erhalten und es exakter zu berechnen.

Die Kommunikation mit anderen Agenten wurde bisher in der Literatur nur für den Fall der benevolenten Agenten behandelt, d.h. Agenten die ihre Daten absolut ehrlich mit anderen teilen (z.B. (Xiang, 1994), (Biswas et al., 1999b)). In dieser Arbeit dagegen wurde die Motivation zum Betrügen analysiert. Dadurch könnte dem Begriff „Betrügen“ eine berechenbare Semantik gegeben werden. Ausgehend davon wurde ein Modell entwickelt, wie Agenten unter Einbezug dieser Motive und dem beobachteten Verhalten auf der Basis eines probabilistischen Ansatzes Rückschlüsse über möglicherweise verheimlichte Daten machen können.

5.3. Das implementierte Agentenverhalten

Die implementierten Agenten sind *sozial situierte* Agenten in dem Sinne, daß sie sich kontinuierlich im Kontakt mit anderen Agenten befinden. Dabei müssen sie eine Reihe von verschiedenen Entscheidungen treffen. Wie und auf welcher Grundlage die implementierten Agenten diese Entscheidung treffen ist in den folgenden Abschnitten erklärt.

5.3.1. Auswahl des eigenen Spielverhaltens

Jeder Agent ist charakterisiert durch einen Wert, der die Wahrscheinlichkeit seines Altruismus und seiner Ehrlichkeit angibt. Nachdem der Agent für seine Spielteilnahme bezahlt hat, ermittelt er zufällig eine Zahl zwischen null und eins. Ist diese

Zahl kleiner als sein Altruismuswert, spielt der Agent altruistisch. Andernfalls spielt er egoistisch. In diesem Fall bestimmt er eine zweite Zufallszahl. Ist diese größer als sein Ehrlichkeitswert, wird er andere Agenten über die Absicht egoistisch zu spielen betrügen und behaupten er spielt altruistisch.

5.3.2. Auswahl eines Spielpartners

Durch das Protokoll des *Offen Gespielten Gefangenendilemmas* erhält jeder Agent von den Agenten, die sich ihm als Spieler anbieten auch ein Angebot, wie sie vorgeben zu spielen. Aufgrund seines Modells dieser Agenten muß er nun entscheiden, ob es sich lohnt mit diesem Agenten und seinem Spielangebot zu spielen.

Hat der Gegner keine Kooperation angekündigt, so lohnt es sich auch nicht mit ihm zu spielen. Solange noch Agenten da sind, die Kooperieren wollen, lohnt es sich einzuschätzen, wieviel Spielpunkte in einer Interaktion mit ihnen zu erwarten sind. Dazu berechnen die Agenten gemäß Definition 17 die Vertrauenswürdigkeit V dieses Agenten und seines Angebots. *pay-off* sei eine Funktion, die die Einträge der Ergebnismatrix zurückgibt unter Berücksichtigung der Tatsache, daß die Teilnahme am Spiel einen Punkt kostet. Da der Agent seinen Spielzug schon angekündigt hat, weiß er zum Zeitpunkt der Partnerauswahl schon, wie er sich entschieden hat zu spielen. Dieser Zug sei mit der Variablen Z beschrieben. Gewichtet man nun gemäß der Bayes'sche Regel das Spielergebnis für den Fall der Vertrauenswürdigkeit bzw. des Gegenereignisses mit den entsprechenden Wahrscheinlichkeiten, so erhält man für den Erwartungswert μ:

$$\mu = V \cdot payoff(Z,C) + (1-V) \cdot payoff(Z,D)$$

Die Nichtteilnahme hat den Punktwert von *0*. Der erste Summand bezeichnet den zu erwartenden Punktgewinn wenn der Gegner sich an sein Spielangebot der Kooperation hält, der zweite den Punktgewinn, wenn der Gegner gelogen hat. Betrachten wir nun den erwarteten Gewinn, für den Fall daß sich der Entscheider für Kooperation entschlossen hat, also $Z=C$ ist. Dies bedeutet für μ*:*

$$\mu = V \cdot payoff(C,C) + (1-V) \cdot payoff(C,D)$$

Damit es sich lohnt zu spielen, muß der erwartete Punktgewinn größer als 0 sein:

$$V \cdot payoff(C,C) + (1-V) \cdot payoff(C,D) > 0$$

$$V \cdot payoff(C,C) - V \cdot payoff(C,D) > -payoff(C,D)$$

$$V \cdot (payoff(C,C) - payoff(C,D)) > -payoff(C,D)$$

$$V > \frac{-payoff(C,D)}{payoff(C,C) - payoff(C,D)}$$

Das beste zu erwartende Ergebnis ergibt sich also bei einer Teilnahme wenn die Vertrauenswürdigkeit des Agenten größer als dieser matrixspezifische Wert ist, andernfalls bei Nichtteilnahme. Dieses Verfahren läßt sich damit direkt auf andere Matrizen übertragen. Die Entscheidung ist nur abhängig von der Vertrauenswürdigkeit des anderen Agenten und der Ergebnismatrix. Analog kann die rationale Entscheidung berechnet werden, wenn der Agent sich für Nicht-Kooperation

entschieden hat. Das weitere Agentenverhalten unterscheidet sich nun, je nachdem welche Rolle der Agent in dieser Spielsituation hat.

---

procedure *akzeptiereDenVertrauenswürdigstenAnbieterAusDerMenge (N)*
begin
$K = \{a \mid a \text{ ist ein Agent der mindestens einmal beobachtet wurde}\}$
befrageZeugenÜber$(K \cup N)$
foreach *Zeugenaussage z* **do**
füge z zum TrustNet hinzu
$q = \arg\max_{a \in N}(\ Vertrauenswürdigkeit(a)\ )$
akzeptiere Spiel mit q
end

Abbildung 20: Algorithmus für den Manager zur Auswahl eines Spielpartners

Der Manager

Ist der Agent in der Rolle des Managers, so hat er eine Menge von Spielangeboten. Aus dieser Menge kann er sich einen Mitspieler aussuchen (dies entspricht dem Argument N in Abbildung 20). Er ist außerdem in der Lage, nochmals andere Agenten nach ihrer Meinung über die Anbieter zu befragen (siehe weiter unten). Hat er deren Aussagen erhalten kann er den erwarteten Punktgewinn für alle Agenten bestimmen. Er wählt dann den Agenten aus, bei dem dieser Punktgewinn maximal ist.

---

procedure *evaluiereManagerAlsSpielpartner(m)*
begin
A_X = *Altruismus des Agenten selbst*
W = *Vertrauenswürdigkeit von m*
if *(*$W+0.05\ W > A_X$ *)*
stimme dem Spiel mit m zu
$K = \{a \mid a \text{ ist ein Agent der mindestens einmal beobachtet wurde}\}$
befrageZeugenÜber$(K \cup \{m\})$
else
lehne Spiel mit m ab
end

Abbildung 21: Algorithmus für den Anbieter zur Auswahl eines Spielpartners

Der Anbieter

In dieser Rolle ist es schwieriger sich für oder gegen ein Angebot zu entscheiden, da der Agent auf ein Spielangebot von einem Manager antwortet und nicht weiß, welche anderen Manager ihm noch Angebote machen werden. Er muß also eine Heuristik wählen. Die implementierte Heuristik besteht darin, daß ein Anbieter die Wahrscheinlichkeit mit der der Manager sich an sein Versprechen hält, berechnet. Er stimmt dem Angebot eines Managers zu, wenn sie größer oder gleich der Wahrscheinlichkeit ist, mit der er selbst kooperiert. Um Rundungsfehler und anfängliche

Ungenauigkeiten zu entgegnen wurde die Schranke in der Implementierung um fünf Prozent tiefer als der Altruismuswert angesetzt (siehe Abbildung 21). Diese Heuristik wird in der Analyse sowohl von der Experimentalgruppe (der Agenten mit `TrustNet`) als auch von der Kontrollgruppe (Agenten ohne `TrustNet`) genutzt.

5.3.3. Zeugen befragen

Gemäß dem Protokoll aus Abschnitt 4.1 haben die Agenten genau einmal pro Runde die Möglichkeit andere Agenten nach deren Beobachtungen zu fragen. Dies findet im Falle des Managers direkt nach dem Erhalt der Spielangebote und im Falle des von ihm akzeptierten Anbieters direkt nach dem Akzeptieren der Paarung statt. Befindet sich nun ein Agent in der Situation, daß er andere befragen darf, stellt sich zunächst die Frage, wen er fragen soll (siehe auch Abbildung 22). Dazu stellt das `TrustNet` eine Funktion zur Verfügung, die die Agenten nach ihrer *(subjektiv ermittelten)* Ehrlichkeit sortiert. Da je nach Experimentkonfiguration eine verschiedene Anzahl von Anfragen an Zeugen erlaubt ist, kann dieser Funktion ein Parameter *n* übergeben werden. Der Rückgabewert der Funktion sind dann die *n* ehrlichsten Agenten. Danach muß entschieden werden, nach wem gefragt werden soll. Um die Implementierung einfach zu halten, findet eine Neuberechnung des ganzen Netzes statt. Der Agent fragt alle Zeugen nach Beobachtungen über alle Agenten deren Identifikation er kennt.

---

procedure *befrageZeugenÜber(q)*
  begin
    $Z = \{ z \mid z \text{ ist unter den n ehrlichsten Zeugen} \}$
    foreach $a \in Z$ **do**
      befrage(a, q)
  end

Abbildung 22: Algorithmus für das Befragen von Zeugen über eine Menge *M* von Agenten

5.3.4. Zeugenaussage machen

Erhält ein Agent die Anfrage eines anderen zu einem Agenten *Q*, so gibt ihm sein `TrustNet` eine Liste aller Beobachtungen die er gesammelt hat. Auf diese wendet er Beobachtung für Beobachtung die *Betrügen*-Funktion mit seinem Ehrlichkeitswert an. Das Ergebnis schickt er als Nachricht an den anfragenden Agenten zurück.

5.4. Technische Realisierung

Im folgenden wird das Testbett beschrieben, in welchem die Entwicklungsumgebung, die Agenten und das `TrustNet` implementiert wurden. Dabei wird auf einige Aspekte der Effizienz eingegangen. Die Dokumentation des Programmes für den Ansatz und seine Experimentalumgebung findet sich in meiner Diplomarbeit.

5.4.1. Social Interaction Framework

Als Testbett wurde das am DFKI in Saarbrücken entwickelte *Social Interaction Framework* (SIF) benutzt. SIF stellt Werkzeuge für die Implementierungen von Multi-Agenten Systemen zur Verfügung (Schillo et al., 1999a). Es benutzt dabei das *Effektor-Medium-Sensor-Modell* (Lind, 1998), welches die Modellierung von situierten, autonomen Agenten vereinfacht und wichtige Grundfunktionalität für die Kommunikation von Agenten zur Verfügung stellt.

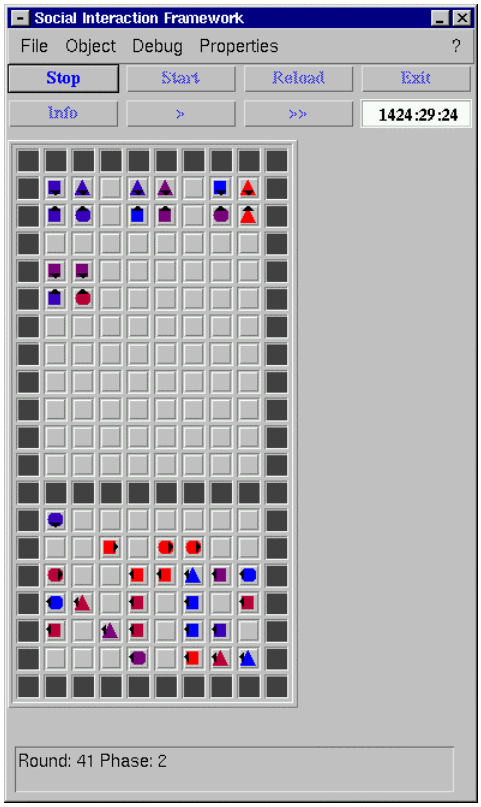

Abbildung 23: Eine Szene während eines Offen Gespielten Gefangenendilemmas

SIF ist in Java™ implementiert und besteht unter anderem aus einer zentralen Einheit (*world server*), die für die Steuerung der Simulation zuständig ist. Das heißt, daß der world server die Gesetzmäßigkeiten (Regeln, Naturgesetze) in der Simulation realisiert. In den im folgenden analysierten Experimenten übernimmt er die Spielleitung, ist also für die Einhaltung der Spielphasen und -regeln zuständig und verteilt die Ressourcen.

In Abbildung 23 ist eine typische Szene der in Abschnitt 4.1 beschriebenen Experimentalumgebung dargestellt. Das Spiel befindet sich in Phase zwei, also der Partnerauswahl. Das Spielfeld ist in zwei Räume geteilt. Die Agenten im oberen Bildbereich haben schon je einen Spielpartner gefunden. Die Spielpaare sind so aufgestellt, daß der eine Spielpartner oberhalb des anderen positioniert ist. Die Spielpaare sind gemäß der Anzahl der anderen Spieler, die sie während eines Spiels beobachten dürfen, in Gruppen angeordnet. In dieser Szene handelt es sich um Gruppen zu je zwei Paaren. Im unteren Bildbereich befindet sich eine Reihe von

Agenten, die noch keinen Interaktionspartner gefunden hat. Der Agent in der linken oberen Ecke in diesem Bereich ist der momentane Manager. Die anderen Agenten im unteren Bildbereich sind Anbieter. Findet der Manager keinen Mitspieler, so muß er den Platz für einen anderen Agenten frei machen. Dieser nimmt seine Position ein und bietet sich als Spielpartner an.

5.4.2. Effizienz

Aus Effizienzgründen ist das `TrustNet` so implementiert, daß es nur dann Knoten neu berechnet, wenn eine Anfrage gestellt wird und seit der letzen Berechnung keine für diesen Knoten relevanten Daten hinzugekommen sind *(lazy evaluation)*. Da die Anfragen gebündelt nach dem Einfügen eigener Beobachtungen und der durch Kommunikation erhaltenen Werte erfolgt, verbessert dies das Laufzeitverhalten. Somit bleibt zwar die theoretische Komplexität der Zugriffsfunktionen erhalten, durch die sparsame Evaluation findet die sehr teure Neuberechnung aber nur selten statt.

Kapitel 6
Evaluation

„There is nothing more necessary to the man of science than its history and the logic of discovery... The way error is detected, the use of hypothesis, of imagination, the mode of testing."

Lord Acton

In diesem Kapitel evaluieren wir den in dieser Arbeit vorgestellte Formalismus empirisch. Für diese Evaluation wurden mit der im letzten Kapitel vorgestellten Experimentalumgebung acht Serien von Simulationen durchgeführt. Jede Serie wurde (zur zuverlässigen statistischen Analyse) vielfach durchgeführt, die erhaltenen Werte wurden über alle Simulationen einer Serie gemittelt. Dabei wurden in etwa 29.000 Stunden Rechenzeit circa 450 Mb Daten gesammelt. Diese wurden dann auf die hier vorgestellten 50 Spielrunden begrenzt und mit Excel aufgearbeitet. Die Daten wurden damit auf 9 Tabellen à 4 Kb reduziert. Sie werden im letzten Abschnitt dieses Kapitels graphisch dargestellt (die numerischen Werte und wie sie reproduziert werden können ist in der gleichlautenden Diplomarbeit beschrieben). Zunächst beschreiben wir die Kriterien, die evaluiert wurden. Im zweiten Abschnitt stellen wir die Variablen vor, die in den Untersuchungen variiert wurden. In Abschnitt 3 analysieren wir den Effekt der einzelnen Variablen auf die Performanz der Agenten im Detail. Eine Auswahl dieser Ergebnisse wurde auch im Workshop „Learning From and About Other Agents" der *International Joint Conference on Artificial Intelligence 1999* in Stockholm präsentiert (Schillo und Funk, 1999).

6.1. Kriterien

Um den Erfolg (bzw. die *Performanz)* des vorgestellten Formalismus beurteilen zu können, haben wir zwei Kriterien genutzt: Zum einen habe wir untersucht, wie gut die Modelle des Verhaltens anderer approximiert wurden, zum anderen haben wir die Performanz der Agenten in Form von erzielten Punkten gemessen.

6.1.1. Wie gut ist das Modell des Verhaltens anderer

Das erste Kriterium nach dem die Agenten und ihre Strategie bewertet wurden ist die Güte ihrer Modelle des Altruismus der anderen Agenten in der Gesellschaft. Dazu

wurde nach jeder Runde für jeden Agenten *X* berechnet, wie groß die Abweichung seines Modells von jedem anderen Agenten *A* von dessen tatsächlicher Altruismuskonfiguration ist. Wir haben hier das übliche Verfahren der Standardabweichung zur Fehlerberechnung zugrundegelegt[1] (siehe z. B. (Bronstein und Semendjajew 1991)):

$$\sigma_X = \sqrt{\sum_A (Modell_X(A) - Konfiguration(A))^2}$$

Aufgrund der Charakteristik dieser Formel ist das Maß des Fehlers für jeden einzelnen Summanden größer null, unabhängig davon, ob der Altruismuswert zu hoch oder zu niedrig eingeschätzt wurde. Die Fehler für jedes einzelne Modell summieren sich also.

6.1.2. Wieviel Punkte hat der Agent erreicht

Das zweite Kriterium ist, wieviel Punkte ein Agent im Laufe des Spiels erreicht hat. Ein Agent beginnt mit zwanzig Punkten in Runde eins. Eine Nichtteilnahme ist kostenlos, die Teilnahme kostet einen Punkt. Am Ende eines Spiels bekommt ein Agent die Punkte gemäß der oben beschriebenen Ergebnismatrix ausgezahlt (siehe Seite 48).

6.2. Variablen

Für die Analyse haben wir acht Simulationsreihen durchgeführt. Jede Simulationsreihe zeichnet sich durch eine spezifische Konfiguration von Variablen aus. Diese Variablen werden hier kurz beschrieben. Eine eingehende Diskussion befindet sich im Abschnitt 6.3, in dem auch die zur jeweiligen Konfiguration gehörigen Ergebnisse dargestellt sind.

6.2.1. Altruismus und Ehrlichkeit

In allen Simulationen haben wir die Auswirkung des Agentenverhaltens auf die Performanz gemessen. Diese Agentenverhalten wird durch unterschiedliche Wahrscheinlichkeiten für egoistisches/altruistisches und ehrliches/betrügerisches Verhalten charakterisiert. Für zwanzig verschiedene Kombinationen dieser Werte wurde die Benutzung von Vertrauen evaluiert.

6.2.2. Mit oder ohne Vertrauen in Zeugenaussagen

Außerdem haben wir den Unterschiede in den beiden Kriterien Modellqualität und Punktergebnis gemessen, je nachdem ob ein Agent die in dieser Arbeit vorgeschlagene Fähigkeit des Vertrauens in Zeugen benutzt hat oder nicht. Um die

---

[1] Die Standardabweichung bezeichnet den Standardfehler des Mittelwertes. In unserer Untersuchung bedeutet dies: je kleiner der Standardfehler σ, umso weniger weit liegen die Modelle von den wirklichen Konfigurationen entfernt. Das Optimum wäre ein Standardfehler von 0.

dafür notwendigen Daten zu erhalten wurden alle Simulationsreihen mit einer Gruppe von Agenten mit und einer Gruppe ohne `TrustNet` durchgeführt (Experimentalgruppe und Kontrollgruppe).

6.2.3. Veränderte Zusammensetzung der Gesellschaft

Diese Untersuchung betrifft die soziale Kompetenz der Gesellschaft: Wie wirkt sich die Benutzung des `TrustNet` aus, wenn es alle Agenten benutzen im Vergleich dazu, daß Agenten ohne diese Fähigkeit in der Gesellschaft sind.

6.2.4. Veränderte Einschränkung der Kommunikation

Hier wird untersucht, wie sich die Veränderung der Menge der erlaubten Anfragen an andere Agenten pro Spielrunde auf die Nützlichkeit des `TrustNet` auswirkt. Diese Variable soll untersuchen, ob das `TrustNet` auch dann nützlich ist, wenn aufgrund der hohen Kosten für Kommunikation nur von wenigen Agenten Daten angefordert werden können. Es soll gezeigt werden in welcher Größenordnung die Kommunikation stattfinden muß um ein gutes Modell der anderen Agenten zu erhalten.

6.2.5. Veränderte Einschränkung der Beobachtung

Dieser Aspekt soll die Frage klären, ob sich die Veränderung der Menge der erlaubten Beobachtungen pro Spielrunde auf die Nützlichkeit des `TrustNet` auswirkt. Hierbei spiegelt sich die Tatsache wieder, daß in den Anwendungsszenarien ein gewisser Anteil von Agenten den Spielern (noch) unbekannt ist.

6.3. Analyse der Performanz von Agenten

Die Performanz der Agenten wurde in jeder Simulation über 50 Spielrunden hinweg aufgezeichnet. Jede Simulationsreihe wurde einhundert mal mit der selben Konfiguration durchgeführt[1]. Für die Analyse jeder Reihe wurde über alle ihre Durchläufe gemittelt.. Es zeigte sich während der Sammlung der Daten, daß die vorgestellten Ergebnisse schon nach dreißig Durchläufen relativ stabil waren. Eine Fortsetzung der Simulationsreihen über 100 Durchläufe hinaus erschien in Anbetracht der entstehenden Rechenlast nicht sinnvoll. Die Simulationen liefen über einen Zeitraum von etwa sechs Wochen auf zwanzig Linux PCs (Pentium II mit 300Mhz, 64Mb Speicher)und zwei Sun Sparc Ultra II (CPU mit 300Mhz, 640Mb Speicher).

---

[1] Die Ausnahme bildet die Simulationsreihe *I*. Aufgrund der Berechnungskomplexität ist diese Reihe besonders rechenintensiv (eine Simulation bis Runde 50 dauert auf einem Linux PC mit Pentium 300 und 64 Mbyte Speicher etwa 10 Tage). Aufgrund der entstehenden Systembelastung, haben wir uns für die Reihe *I* auf 50 Simulationen beschränkt. Aus der Varianz der Daten aus den vorliegenden 50 Durchläufen läßt sich schließen, daß weitere 50 Durchläufe keine widersprechenden Ergebnisse liefern würden.

In jeder Simulation bestand die Gesellschaft aus insgesamt 40 Agenten. Mit Ausnahme der Simulationsreihe *N* unterschied sich in jeder Reihe jeder Agent von allen anderen durch seine Konfiguration (in Simulation *N* gab es von jeder Konfiguration zwei Agenten, da die Kontrollgruppe entfiel). Die Agenten charakterisieren sich, wie in Kapitel 4 dargestellt durch zwei Werte: Ehrlichkeit und Altruismus. Für Altruismus wurden die fünf Werte 0,01; 0,25; 0,50; 0,75 und 0,99 gewählt, für Ehrlichkeit die vier Werte 0,01; 0,33; 0,66 und 0,99. Dadurch ist gewährleistet, daß keiner der Agenten ein hundertprozentig eindeutiges Verhalten an den Tag legt und die Einschätzung seiner Konfiguration nicht zu einfach ist. Wie in Abbildung 24 dargestellt, setzt sich die Gesellschaft einer Simulation allen möglichen Kombinationen dieser Werte zusammen. Dadurch wird sichergestellt, daß keine für das Ergebnis wesentliche Kombination der Werte ausgelassen wird. Für jede Konfiguration wurden zwei Agenten in die Gesellschaft aufgenommen: Ein Agent ohne `TrustNet` (diese Agenten bilden die sog. Kontrollgruppe) und ein Agent mit `TrustNet` (diese Agenten bilden die sog. Experimentalgruppe). Durch den Vergleich zwischen der Performanz der Kontrollgruppe und der Performanz der Experimentalgruppe läßt sich für jede Simulationsreihe zeigen, ob die Benutzung des `TrustNet` von Nutzen war oder nicht. Da keine zwei Agenten das gleiche Verhalten haben, sind die Gesellschaften der Simulationsreihen also in hohem Maße *heterogen*. Alle Simulationen liefen 50 Spielrunden lang. Nach jeder Runde wurden für alle Agentenkonfigurationen beide Kriterien gemessen.

		Altruismus				
Ehrlichkeit		0,01	0,25	0,50	0,75	0,99
	0,01	1/1[1]	1/1	1/1	1/1	1/1
	0,33	1/1	1/1	1/1	1/1	1/1
	0,66	1/1	1/1	1/1	1/1	1/1
	0,99	1/1	1/1	1/1	1/1	1/1

Abbildung 24: Zusammensetzung der Agentengesellschaft

Aufbau der Analyse: Im nächsten Abschnitt wird die Performanz von Agenten im *Offen Gespielten Gefangenendilemma mit Partnerwahl* in absoluten Werten für Punkte und Modellqualität beschrieben. In den darauffolgenden Abschnitten wird zur deutlicheren Darstellung nur das Verhältnis zwischen Kontroll- und Experimentalgruppe beschrieben. Wie in Abschnitt 6.2 dargestellt, werden fünf Variablen untersucht. Diese Variablen spannen einen „Konfigurationsraum" mit fünf Dimensionen auf. Drei davon sind in Abbildung 25 dargestellt: die Zusammensetzung der Gesellschaft, der Anteil der Gesellschaft, mit der die Agenten kommunizieren dürfen und der Anteil, den sie beobachten dürfen. Um die Komplexität zu reduzieren, wurde auf die vollständige Untersuchung dieses Raumes verzichtet. Stattdessen wurde ausgehend von der Konfiguration *G* jede Variable verändert, während die jeweils anderen Variablen festgehalten wurden. Die vierte Dimension

[1] Je ein Agent mit und ein Agent ohne `TrustNet`.

(Kontrollgruppe vs. Experimentalgruppe) ist in Abschnitt 6.3.2 erklärt und wird für alle Simulationen betrachtet. Die fünfte Dimension (das Agentenverhalten, siehe 6.2.1) wurde ebenfalls in allen Reihen betrachtet. In einigen Untersuchungen gibt es in dieser Dimension allerdings keine signifikanten Unterschiede. Deshalb wird sie in diesen Untersuchungen dann nicht explizit dargestellt.

Die Untersuchung Kontroll- vs. Experimentalgruppe in Abschnitt 6.3.2 führt den Beweis, daß sich die Performanz der Agenten verbessert, wenn sie das `TrustNet` benutzen. Die weiteren Untersuchungen gehen darauf ein, wie sich diese Performanzzunahme ändert, wenn die Konfiguration der Simulation verändert wird.

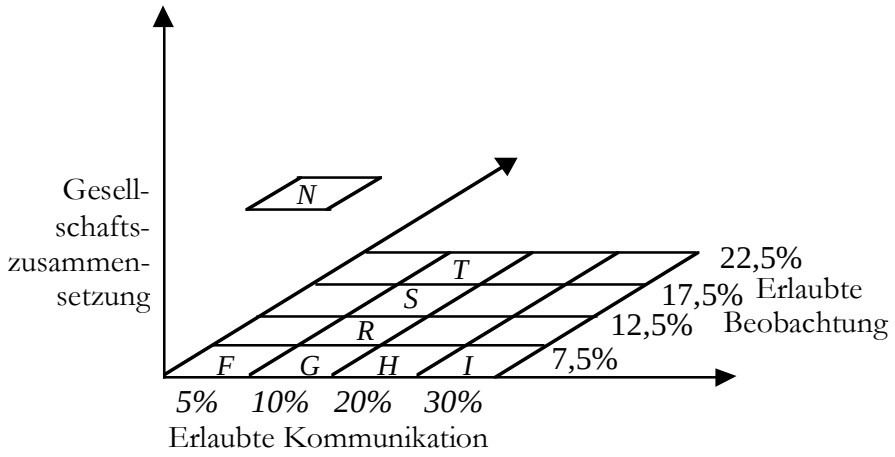

Abbildung 25: Der Zusammenhang zwischen den Simulationsreihen.

Drei der untersuchten Variablen: die Zusammensetzung der Gesellschaft, der Anteil der Gesellschaft, mit der die Agenten kommunizieren dürfen und der Anteil, den sie beobachten dürfen.

6.3.1. Altruismus und Ehrlichkeit

Zunächst geben wir einen Überblick über die generelle Performanz von Agenten im beschriebenen Spiel. Dieser Abschnitt gibt das Spielergebnis von Agenten wieder, die ohne das `TrustNet` auskommen müssen *(einfache Agenten)*. Während der Simulation benutzen sie nur ihre eigenen Beobachtungen. Dies ist der einzige Punkt, in dem sich ihre Berechnungen und ihr Verhalten zu den in Kapitel 4 beschriebenen Agenten unterscheidet. Insbesondere benutzen alle Agenten das Vertrauensmodell von Castelfranchi und Falcone zur Berechnung des erwarteten Nutzens der Kooperation mit Spielpartnern. Für diese Analyse wurden die Daten der Simulationsreihe *G* gewählt. Die Konfiguration der Reihe *G* zeichnet sich dadurch aus, daß sie in den Anwendungsszenarien als realisierbar erscheint und wurde daher als Ausgangskonfiguration gewählt. Die Auswirkungen von Altruismus und Ehrlichkeit der Agenten aus der Experimentalgruppe betrachten wir in jeder der folgenden Abschnitte gesondert.

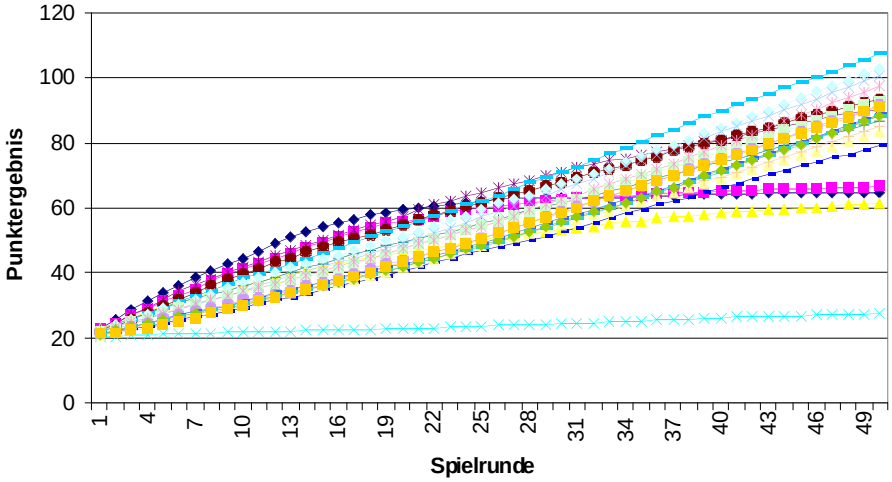

Abbildung 26: Das Punktergebnis aller einfachen Agenten im Überblick

Eine Graphik mit den Punktergebnissen aller Agenten aus Reihe *G*, die das `TrustNet` nicht benutzen, zeigt Abbildung 26. Dieses Diagramm soll bereits einen ersten Eindruck über den ungefähren Verlauf der Punktergebnisse der Agenten geben. Da es aufgrund der vielen Graphen schwer zu interpretieren ist, stellen wir jeweils die Agenten mit dem selben Ehrlichkeitswert in einem eigenen Diagramm vor.

Für jeden der vier Ehrlichkeitswerte in der Agentengesellschaft steht eine der Abbildungen 27 bis 30. In jeder Abbildung sind fünf verschiedenen Graphen für die fünf Konfigurationen mit dem jeweiligen Ehrlichkeitswert und den verschiedenen Altruismuswerten dargestellt. Zusammen ergibt dies die zwanzig verschiedenen Agentenkonfigurationen ohne `TrustNet`. Wir interpretieren nun die abgebildeten Punktergebnisse. Wie schon gesagt, befinden sich die zugrundeliegenden Daten auch in der gleichlautenden Diplomarbeit.

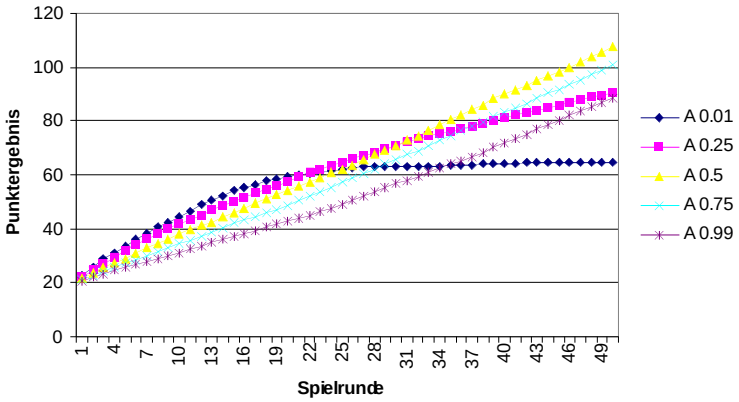

Abbildung 27: Punktergebnis einfacher Agenten mit Ehrlichkeitswert 0,01

Hierbei steht *A* für den Altruismuswert eines Agenten. Jeder Graph stellt also das Mittel des Punktergebnisses von fünf verschiedenen Agenten dar.

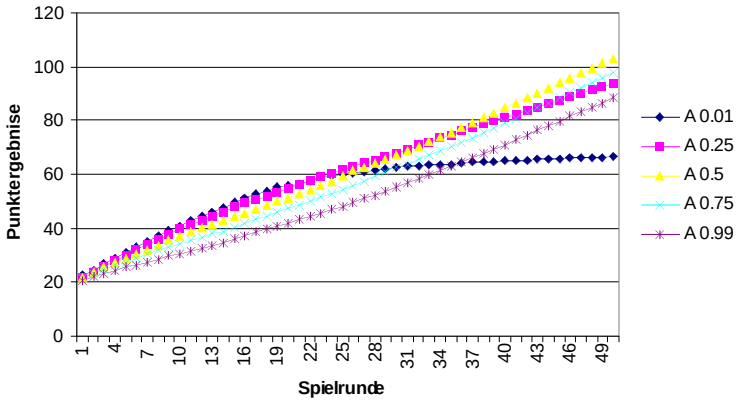

Abbildung 28: Punktergebnis einfacher Agenten mit Ehrlichkeitswert 0,33

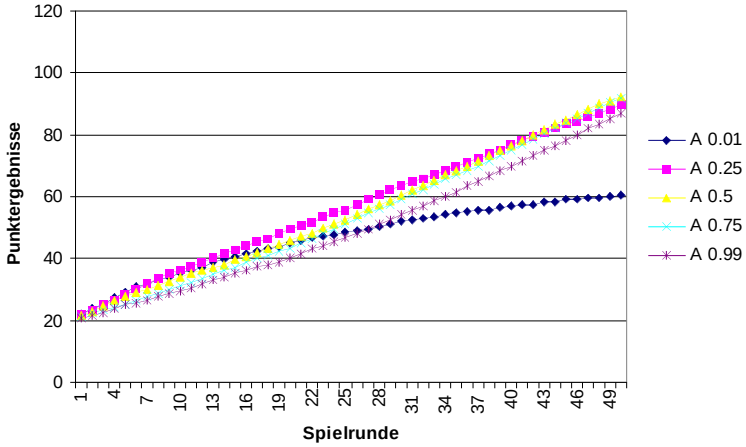

Abbildung 29: Punktergebnis einfacher Agenten mit Ehrlichkeitswert 0,66

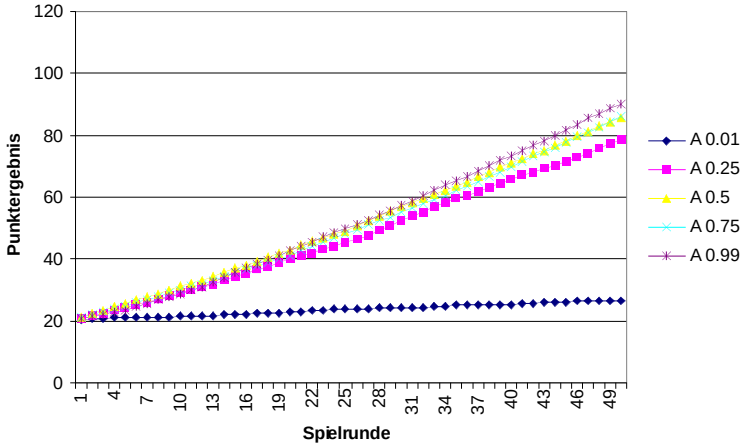

Abbildung 30: Punktergebnis einfacher Agenten mit Ehrlichkeitswert 0,99

Die vier altruistischen Agenten (Wert 0,99) weisen anfänglich relativ geringe Punktgewinne auf. Nachdem ihr Punktekonto zu steigen beginnt, wächst es fast linear. Daraus läßt sich ablesen, daß sie nach einer gewissen Zeit nur ebenso altruistische Spielpartner auswählen und nicht mehr ausgenutzt werden. Da sie sich trotz anfänglich sehr schlechter Modelle gegenüber anderen fast ausschließlich kooperativ verhalten, war diese Entwicklung zu erwarten. Bis Runde 50 ist ihr Verhalten aber schlechter, als das der Agenten mit Altruismuswerten 0,50 und 0,75. Einzig der ehrliche und altruistische Agent (0,99; 0,99) schlägt ab Runde 15 alle anderen, die genauso ehrlich sind wie er. Die Punkteperformanz des egoistischen Agenten (Altruismuswert 0,01) wächst am Anfang sehr stark, da er die Unwissenheit der anderen ausnutzen kann und flacht dann auf einen fast konstanten Verlauf ab. Er schneidet um so schlechter ab, je ehrlicher er ist. Im Falle, daß er in 99 Prozent der Fälle ehrlich ist, entfällt sogar der anfängliche Anstieg. Ein Verhalten, das auch zu erwarten war, da er ja in diesem Fall ehrlich bzgl. seiner Intentionen ist. Interessant ist, daß sich in den Graphen zeigt, daß der altruistische Agent um so besser wird, je schlechter der egoistische ist.

In allen Diagrammen gibt es einen Punkt, ab dem die altruistischeren Agenten die egoistischeren überholen. Dies liegt daran, daß mit der Zeit die Modelle des Verhaltens der anderen besser werden. Die besseren Modelle liefern ein an das Verhalten der Interaktionspartner angepasstes Verhalten: Agenten suchen oder vermeiden Spielpartner aufgrund deren Konfiguration, die sie mit ihren besseren Modell angenähert haben.

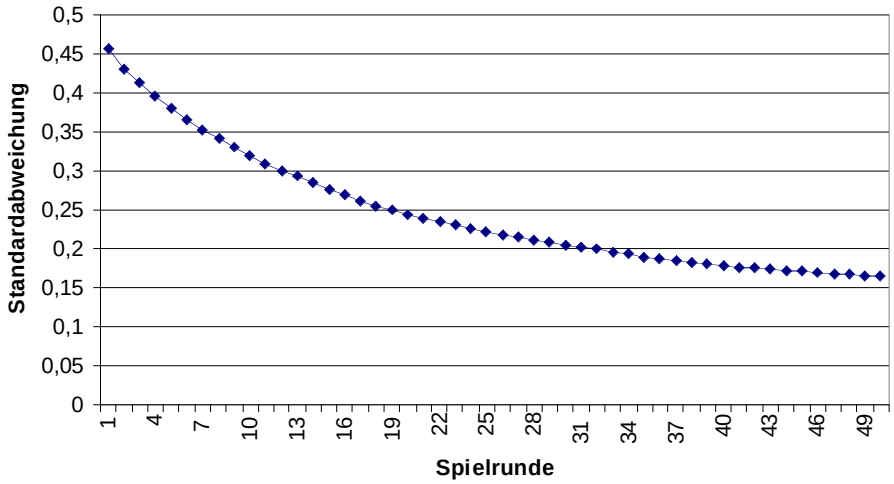

Abbildung 31: Die Modellqualität einfacher Agenten im Spielverlauf

Die Qualität dieses Modells wird durch die Berechnung der Standardabweichung von dem realen Wert berechnet (das σ_X eines Agenten *X*, siehe oben). Ein Wert von 0 wäre das Optimum (Modell gleich reale Konfiguration). Da die Konfiguration eines Agenten keine Auswirkungen auf die Qualität seiner Modelle hat, wird hier über alle zwanzig Agenten gemittelt. Diese Daten sind in Abbildung 31 dargestellt. Wie zu erwarten, sinkt der durchschnittliche Fehler mit zunehmender Spielerfahrung. Die Kurve hat eine schwache Hyperbelform. Daraus ergibt sich, daß die Modelle am Anfang sehr rasch an Güte gewinnen, danach eine weitere Verbesserung jedoch sehr

lange dauert. In Zahlen ausgedrückt: um die selbe Verbesserung des Modells wie zwischen Runde eins und zwölf zu erreichen brauchen sie nach Runde zwölf länger als 38 Runden.

Der nächste Abschnitt zeigt die Werte für die Agenten mit `TrustNet` und deren Verbesserung ihrer Performanz gegenüber dieser Gruppe.

6.3.2. Mit oder ohne Vertrauen in Zeugenaussagen

Dieser Abschnitt der Analyse ist von zentraler Bedeutung für diese Arbeit. Er zeigt, wie sich die Performanz von Agenten ändert, wenn sie den vorgeschlagenen Algorithmus und seine Implementierung, das `TrustNet`, benutzen. Dazu stellen wir im folgenden die prozentuale Veränderung der Performanz der Experimentalgruppe im Vergleich zur Kontrollgruppe dar. Wie im vorhergehenden Abschnitt wollen wir die Agenten, aufgeteilt nach Ehrlichkeitswerten betrachten. Positive Werte in den Diagrammen von Abbildung 32 bis Abbildung 35 stellen den Performanzgewinn der Agenten mit `TrustNet` dar.

Das erste Ergebnis dieser Untersuchung ist, daß fast alle Agentenkonfigurationen von der Verwendung des `TrustNet` einen statistisch signifikanten Gewinn haben. Lediglich die egoistischen Konfigurationen (Altruismuswert 0,01) haben keinen Gewinn, ihre Werte liegen unter fünf Prozent im positiven bzw. negativen Bereich und haben nur geringe (statistische) Aussagekraft. Außerdem zeigt sich, daß alle Diagramme eine ähnliche Form haben. Die eigene Ehrlichkeit eines Agenten hat also wenig bis keinen Einfluß auf den Nutzen der durch Vertrauen in andere erzielt wird. Insbesondere hoch sind die Punktgewinne für die Agenten mit den Altruismuswerten 0.99, 0.75 und 0.50. Im Mittel liegt deren Gewinn nach einer Phase der Datensammlung von zehn Interaktionen bei 15 bis 20 Prozent. Außerdem ist bemerkenswert, daß die Agenten umso mehr Gewinn vom `TrustNet` haben, je höher ihr Altruismuswert ist.

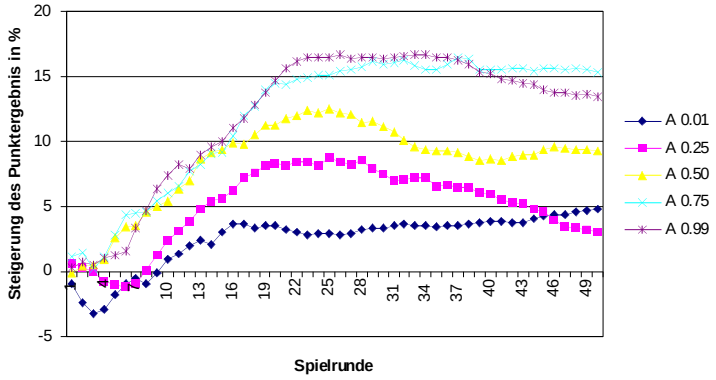

Abbildung 32: Der prozentuale Performanzanstieg der Experimentalgruppe mit Ehrlichkeitswert 0,01 im Vergleich zur gleichen Gruppe ohne `TrustNet`.

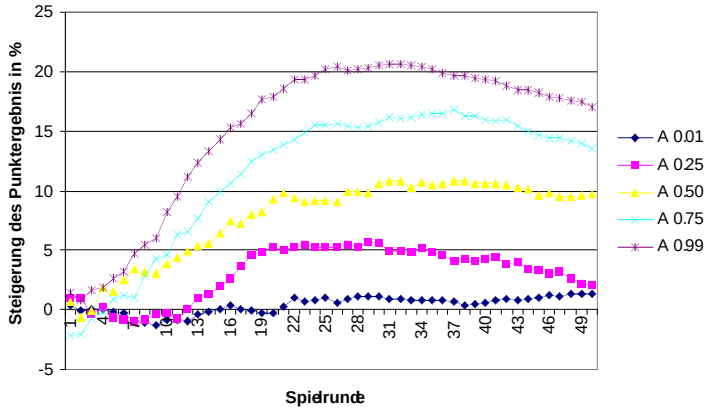

Abbildung 33: Der prozentuale Performanzanstieg der Experimentalgruppe mit Ehrlichkeitswert 0,33 im Vergleich zur gleichen Gruppe ohne `TrustNet`.

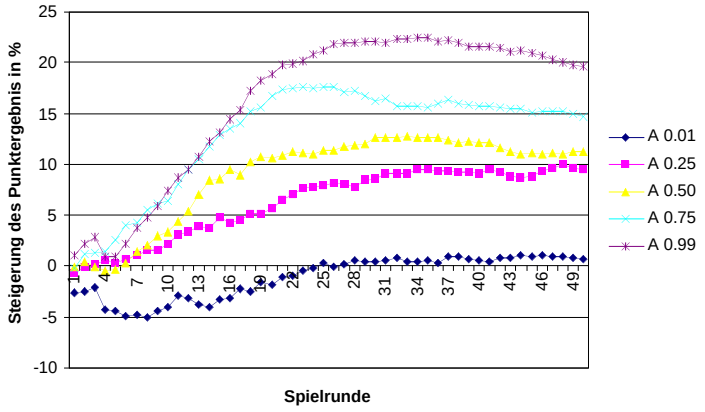

Abbildung 34: Der prozentuale Performanzanstieg der Experimentalgruppe mit Ehrlichkeitswert 0,66 im Vergleich zur gleichen Gruppe ohne `TrustNet`.

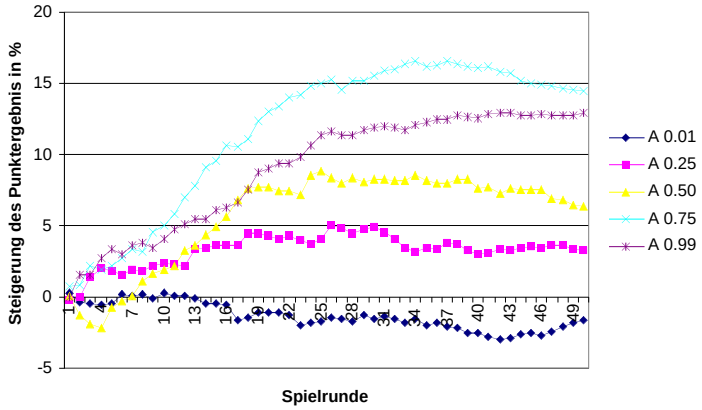

Abbildung 35: Der prozentuale Performanzanstieg der Experimentalgruppe mit Ehrlichkeitswert 0,99 im Vergleich zur gleichen Gruppe ohne `TrustNet`.

Da sich die eigene Konfiguration der Agenten nicht auf die Modellbildung auswirkt, haben wir auf eine Aufschlüsselung der Modelle nach Ehrlichkeitswert verzichtet. Stattdessen zeigt Abbildung 36 die Verringerung des Modellfehlers der gesamten Experimentalgruppe gegenüber der ganzen Kontrollgruppe. Es zeigt sich, daß die Experimentalgruppe (die sich nur in der Verwendung des `TrustNet` von der anderen Gruppe unterscheidet) eine enorme Verbesserung ihrer Modelle erreicht hat. Nach nur sechs Runden ist die Verbesserung deutlich über fünf Prozent. Nach weiteren sechs Runden liegt sie bereits bei zwanzig Prozent, ein Niveau, daß über zwanzig Runden gehalten wird. Danach sinkt die Verbesserung durch das `TrustNet` allmählich, bleibt bis Runde 50 aber über zehn Prozent. Insbesondere im Bereich, in dem noch wenig Beobachtungen anderer gemacht werden konnten, bedeutet das `TrustNet` folglich eine wesentliche Verbesserung der Performanz.

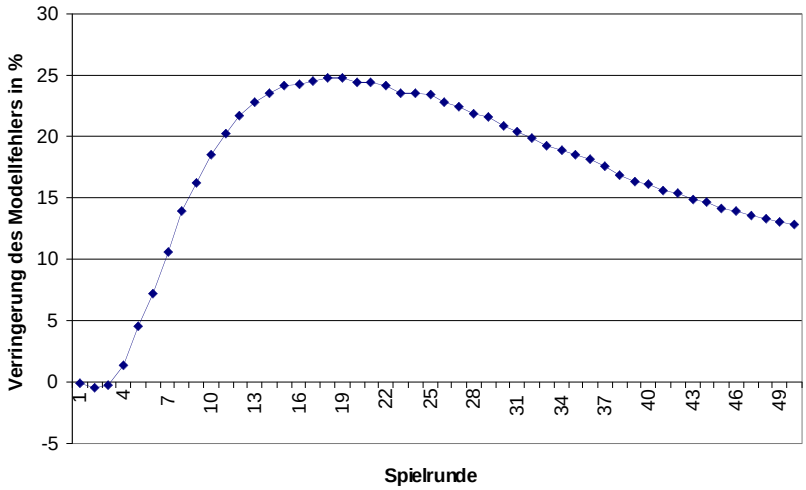

Abbildung 36: Die Verbesserung der Modellqualität durch Verwendung des `TrustNet`

6.3.3. Veränderte Zusammensetzung der Gesellschaft

Wir untersuchen in diesem Abschnitt eine Veränderung der Zusammensetzung der Gesellschaft. Wie schon erwähnt, enthält die Simulationsreihe *G* eine Kontrollgruppe, die das `TrustNet` nicht benutzt. Aufgrund dieser Eigenschaft könnte man davon sprechen, daß diese Gruppe dadurch weniger soziale Kompetenz besitzt. Wir ersetzen diese Gruppe in der Simulationsreihe *N* durch eine zweite Experimentalgruppe, um damit die soziale Kompetenz in der Gesellschaft insgesamt zu erhöhen, aber gleichzeitig die Gesamtzahl der Agenten beizubehalten. Im Vergleich zu Abbildung 24 (die Zusammensetzung der Gesellschaft in Konfiguration *G*) zeigt Abbildung 37 die Zusammensetzung in der Konfiguration *N*. Es sind nun von jeder Konfiguration zwei Agenten mit `TrustNet` vorhanden.

Diese Veränderung der Gesellschaft wirkt sich auch auf das Punktergebnis der Agenten aus. Da der Ehrlichkeitswert an dieser Stelle keine wesentlichen Aussagen hinzufügt, wurde auf eine Aufschlüsselung nach Ehrlichkeit verzichtet. Abbildung 38 zeigt also die Mittelwerte für alle Agenten mit dem angegebenen Altruismuswert. Die

dort eingezeichneten Graphen veranschaulichen die Veränderung des Punktergebnisses der Experimentalgruppe in Konfiguration *N* gegenüber der im letzten Abschnitt vorgestellten Konfiguration *G*. Ein negativer Wert bedeutet, daß die entsprechenden Agenten weniger Punkte in Konfiguration *N* erhalten haben als an der selben Stelle in Konfiguration *G*.

		Altruismus				
Ehrlichkeit		0,01	0,25	0,50	0,75	0,99
	0,01	2	2	2	2	2
	0,33	2	2	2	2	2
	0,66	2	2	2	2	2
	0,99	2	2	2	2	2

Abbildung 37: Zusammensetzung der Agentengesellschaft in Simulationsreihe *N*

Es zeigt sich, daß die Egoisten der Experimentalgruppe deutlich schlechter abschneiden, während die Altruisten der selben Gruppe ungefähr auf dem selben Punkteniveau bleiben. Daraus läßt sich schließen, daß die Egoisten in der weniger sozial kompetenten Gesellschaft vermehrt die Agenten ohne `TrustNet` ausgenutzt haben. Dies unterstützt die Beobachtung aus dem letzten Abschnitt, daß die altruistischeren Agenten besonders stark vom `TrustNet` profitieren. Dies ist ein weiterer Hinweis für die Verbesserung, die sich durch die Verwendung des `TrustNet` ergibt. Die altruistischen Agenten mit `TrustNet` finden nach wie vor ihre Spielpartner. Ihr Punktekonto verändert sich daher kaum.

Daraus ergeben sich zwei Folgerungen: Erstens sinkt in sozial kompetenteren Gesellschaften unter Verwendung von Vertrauen die Performanz für die Egoisten. Zweitens erklärt sich die, absolut gesehen, schlechtere Performanz der Altruisten mit `TrustNet` gegenüber den Egoisten mit `TrustNet` in den anderen Studien dadurch, daß die Egoisten vermehrt unwissendere Agenten der Kontrollgruppe ausnutzen.

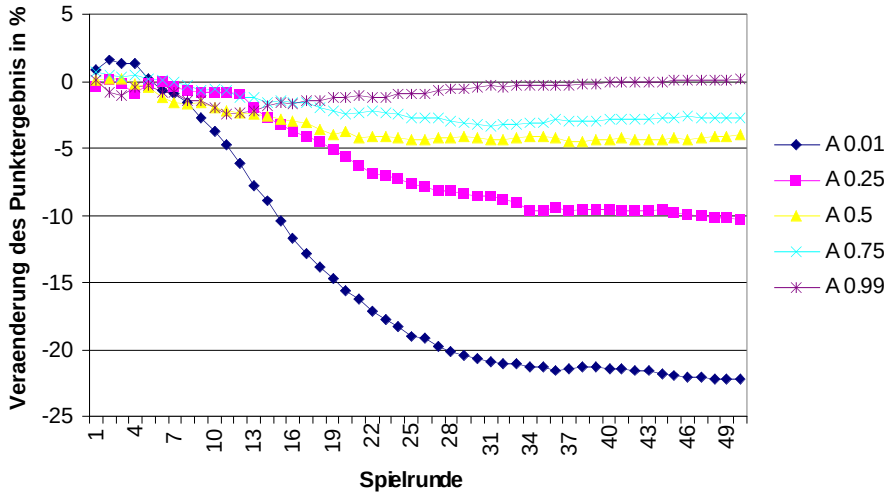

Abbildung 38: Veränderung des Punktergebnisses der Agenten mit `TrustNet` in einer sozial kompetenteren Gesellschaft gegenüber der Anfangskonfiguration

6.3.4. Veränderte Einschränkung der Kommunikation

Während der Phase der Partnerwahl hat der Manager vor der Wahl seines Interaktionspartners die Möglichkeit mit *n* anderen Agenten zu kommunizieren. Ausgehend von der Konfiguration *G*, bei der die Agenten von zehn Prozent der Gesellschaft Daten anfordern durften, wurden drei weitere Simulationsreihen durchgeführt. Eine Reihe mit weniger Kommunikation (*F*) und zwei Reihen mit mehr Kommunikation (*H*, *I*). Eine Evaluation mit noch mehr Kommunikation erschien nicht sinnvoll, da Kommunikationskosten in MAS im allgemeinen sehr hoch sind. Die untersuchten Konfigurationen sehen folgendermaßen aus:

Simulationsreihe	Anzahl der erlaubten Anfragen	Anteil an der Gesellschaft
F	2	5%
G	4	10%
H	8	20%
I	12	30%

Abbildung 39 zeigt die unterschiedlichen Steigerungen der Punktergebnisse der Experimentalgruppen (der Agenten mit `TrustNet`) je nach Simulationsreihe. Dabei ist nur der prozentuale Unterschied zur Kontrollgruppe aufgetragen. Ein Wert von zehn bedeutet hier also eine um zehn Prozent bessere Punktperformanz. Der in Abschnitt 6.3.2 festgestellte grundlegende Kurvenverlauf zeigt sich hier wieder in allen Simulationsreihen. Mit späteren Spielrunden hat die Kontrollgruppe genug Daten gesammelt, daß ihre Performanz so gut ist, daß der Abstand zur Experimentalgruppe abnimmt. Es zeigt sich, daß die Performanz der Agenten mit `TrustNet` steigt, je mehr Kommunikation erlaubt ist. Nach einem Schwanken in den ersten vier Spielrunden[1], gilt für die Werte der Simulationsreihen, daß die Kurven sich verhalten wie F<G<H<I. Das ist in sofern ein positives Ergebnis, als die Performanz des `TrustNet` demnach mit wachsender Information auch bessere Ergebnisse liefert. Gegen Runde 50 flacht die Kurve ab. In explorativen Studien[2] hat sich aber gezeigt, daß dieses Abflachen noch sehr lange über einem Niveau von 5 Prozent bleibt und also der Punktgewinn gegenüber der Kontrollgruppe auch jenseits der Runde 50 andauert.

---

[1] Dieses Schwanken erklärt sich durch die zufällig bestimmte Anfangspositionierung, die wesentlich die Spielpaare in der ersten Runde beeinflußt, da in dieser Runde keiner der Agenten über Daten von anderen besitzt. In diesem Zusammenhang ist es wichtig festzustellen, daß dieses Schwanken der Daten um die X-Achse nach Runde fünf nicht mehr auftritt. Daraus folgt, daß die Benutzung des TrustNet, nicht nur darin besteht, daß es die Performanz verbessert, sondern die Performanz auch stabilisiert.

[2] Deren Ergebnisse sind aufgrund nur weniger Durchläufe nicht statistisch gesichert und werden daher an dieser Stelle nicht weiter diskutiert.

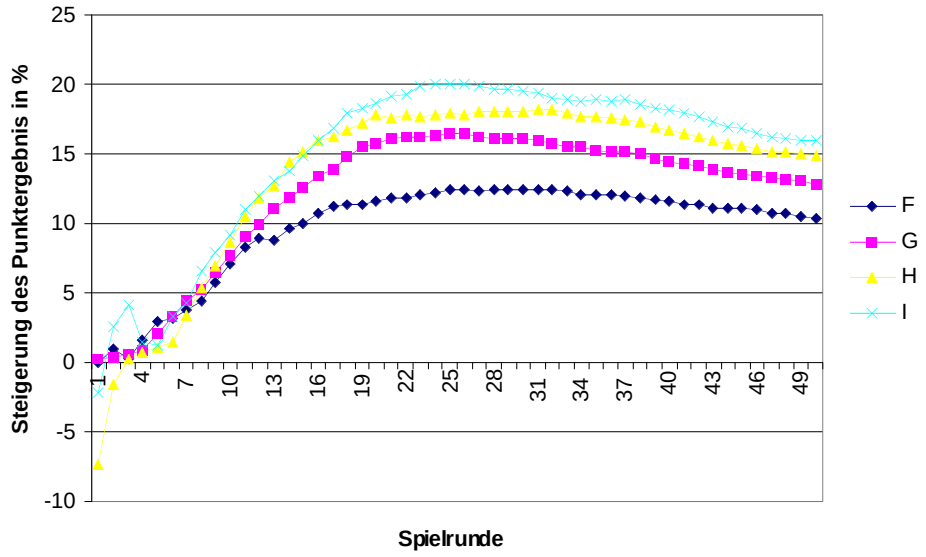

Abbildung 39: Prozentuale Verbesserung des *Punktergebnisses* durch Verwendung des `TrustNet` bei unterschiedlichen Einschränkungen der Kommunikation

Ein wichtiges Ergebnis dieser Betrachtung ist, daß zwar der Punktgewinn zwischen fünf und zehn Prozent Kommunikation relativ groß ist (ein Plus von etwa 5-10 Prozentpunkten), interessanterweise der Gewinn der Konfigurationen *H* und *I* gegenüber *G* aber nicht so stark ist. Dies zeigt ein für die Anwendungsszenarien wichtiges Ergebnis: Es gibt eine relativ *niedrige* Schranke ab der die Kommunikation keine wesentlichen Performanzsteigerungen mehr ergibt. Für das `TrustNet` zeigt sich, daß es schon mit sehr wenig Kommunikation sehr gute Ergebnisse liefert. Dies kann darauf zurückgeführt werden, daß das `TrustNet` nicht nur vertrauenswürdige Interaktionspartner berechnet, sondern auch angibt, welcher der Informanten am Besten befragt werden sollte. Dadurch kommt wohl ein Agent schon mit sehr wenigen Informanten aus.[1]

Die gute Performanz läßt sich wie in den vorigen Abschnitten auf die Entwicklung der Modelle der Agenten zurückführen. Diese ist in Abbildung 40 dargestellt. Daß die Steigerungen der Punktperformanz relativ nahe beieinander liegen, hat seinen Grund darin, daß auch die Maxima der Modellverbesserungen fast identisch sind. Ihre Unterschiede erklären sich allein dadurch, daß die Konfigurationen mit mehr Kommunikation dieses Optimum schneller erreichen (*G* etwa bei Runde 15, *F* erst bei Runde 20). Wie schon für die Punktperformanz festgestellt, liegt die Konfiguration *G* (Kommunikation mit zehn Prozent der Gesellschaft) deutlich vor der Konfiguration *F*, während der Abstand zu *H* und *I* deutlich geringer ausfällt.

[1] Dies ist ein Hinweis darauf, daß in größeren Gesellschaften eine bessere Performanz auch dann erreicht werden kann, wenn der Anteil der Agenten mit denen kommuniziert werden darf, weiter sinkt. Dies liegt daran, daß die absolute Anzahl der Informanten sehr hoch ist.

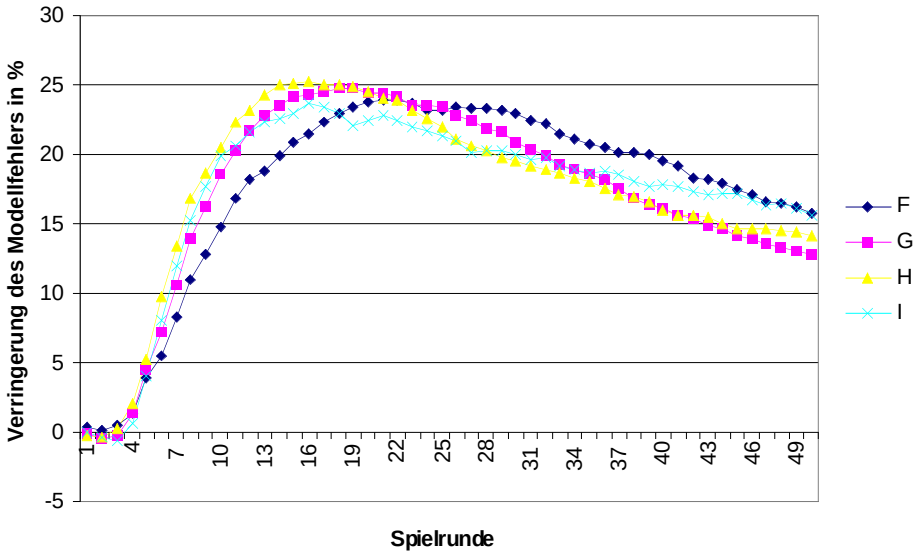

Abbildung 40: Prozentuale Verbesserung der *Modelle* durch Verwendung des `TrustNet` bei unterschiedlichen Einschränkungen der Kommunikation

6.3.5. Veränderte Einschränkung der Beobachtung

Die letzte der untersuchten Variablen ist die Einschränkung der Beobachtung. Hier wird untersucht, wie sich die Verringerung oder Vergrößerung der Anzahl der beobachteten Spieler pro Runde auf das Ergebnis auswirkt. Ausgehend von der Konfiguration *G* (Beobachtung von 7,5 Prozent der Gesellschaft) wurden drei Studien mit jeweils mehr erlaubter Beobachtung pro Runde durchgeführt.

Simulationsreihe	Anteil an der Gesellschaft	Anzahl absolut
G	7,5%	3
R	12,5%	5
S	17,5%	7
T	22,5%	9

Je mehr Beobachtung in dieser Untersuchung erlaubt ist, desto schneller wird eine gute Performanz erreicht (siehe Abbildung 41). Außerdem gilt, daß je mehr Beobachtung pro Runde zur Verfügung steht der maximale Gewinn durch das `TrustNet` sinkt. Ist das Maximum für Konfiguration *G* noch jenseits der 15 Prozent, so liegt es bei Konfiguration *T* nur noch etwas über 10 Prozent. Dies liegt daran, daß bei den Konfigurationen mit mehr Beobachtungen, auch den Agenten der Kontrollgruppe sehr viele Daten zur Verfügung stehen. Durch die hohe Verfügbarkeit von Daten reduziert sich natürlich der Effekt der Kommunikation mit anderen, da deren Daten immer weniger notwendig sind. Hier sei angemerkt, daß das `TrustNet`, trotz der sehr hohen Beobachtungsrate (die für die Anwendungs-

szenarien schon unrealistisch hoch ist) immer noch einen signifikanten Punktgewinn liefert.

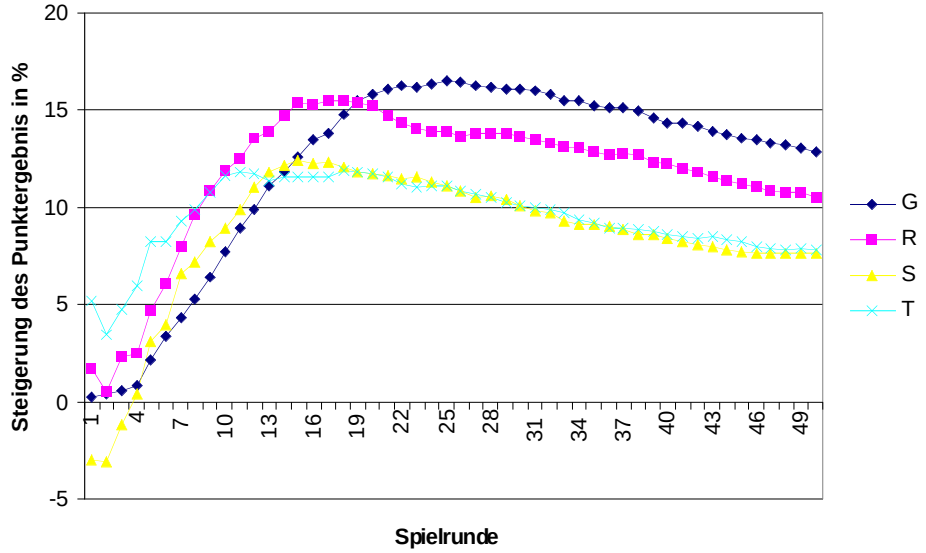

Abbildung 41: Prozentuale Verbesserung des *Punktergebnisses* durch Verwendung des `TrustNet` bei unterschiedlichen Einschränkungen der Beobachtung pro Runde

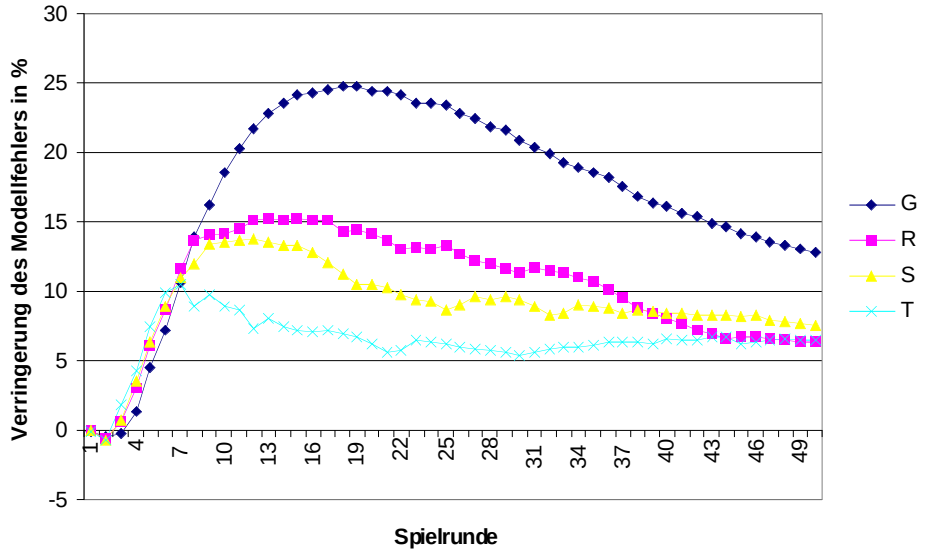

Abbildung 42: Prozentuale Verbesserung der *Modelle* durch Verwendung des `TrustNet` bei unterschiedlichen Einschränkungen der Beobachtung pro Runde

Überraschend in Abbildung 41 ist, daß nicht nur die Maxima der Kurven sich unterscheiden, sondern auch die Geschwindigkeit, mit der die Kurven ansteigen. Es zeigt sich, daß mehr Beobachtung auch dazu führt, daß das `TrustNet` sich schneller auszahlt. Dies weist darauf hin, daß selbst unter günstigen Beobachtungsumständen die Verwendung des `TrustNet` einen (wenn auch kleinen) positiven Einfluß auf den

Punktgewinn hat. Insbesondere in Situationen, in denen schwer abzuschätzen ist, welche Konfiguration vorliegt, ist diese Erkenntnis von hohem Wert, da aus ihr folgt, daß durch die Benutzung des `TrustNet` kein Nachteil entsteht.

Das oben gesagte spiegelt sich auch in der Analyse der Modellqualität wieder. Die Konfigurationen mit hohem Beobachtungsanteil erreichen ihr Gewinnoptimum sehr schnell. Dies liegt daran, daß bei hohem Beobachtungsanteil auch bessere Informationen vorliegen, welche Agenten bessere Zeugen sind. Des weiteren zeigt sich, daß je weniger Beobachtung zur Verfügung steht, das `TrustNet` von hohem Vorteil ist. Erst ab Runde 40 nähern sich die Graphen an die 5 Prozent Grenze an und abgesehen von Schwankungen in den ersten drei Runden sinken sie nie in den negativen Bereich ab. Die Maxima sind nach etwa zehn Interaktionen erreicht und liegen bei etwa zwanzig, fünfzehn, zwölf und acht Prozent, je nach Konfiguration. Gerade in dem für die Anwendungsszenarien wichtigen Bereich zeigt sich also das `TrustNet` erfolgreich. Außerdem zeigt sich, daß der anfänglich erreichte Vorsprung in der Modellqualität auch noch lange anhält. Schließlich verdeutlicht die Graphik, daß in allen vier Konfigurationen das `TrustNet` signifikant bessere Modelle liefert, als sie aufgrund der eigenen Beobachtungen möglich gewesen wären.

Kapitel 7
Ergebnis

"We inhabit a climate of *trust* as we inhabit an atmosphere and notice it as we notice air, only when it becomes scarce or polluted."
A. Baier, 1977.

Im ersten Abschnitt fassen wir zusammen, welche Schlüsse sich aus dieser Arbeit ziehen lassen. Der zweite Abschnitt berichtet über den wissenschaftlichen Beitrag der aus der vorliegenden Formalisierung und ihrer Analyse folgt. Im letzten Abschnitt schließlich geben wir eine Reihe von möglichen Erweiterungen dieser Arbeit an, die sich während der Implementierung und Analyse gestellt haben. Die Form für diese Zusammenstellung folgt den Empfehlungen von Chinneck (1999).

7.1. Schlußfolgerungen

Der besseren Übersicht wegen stellen wir die wichtigsten Schlußfolgerungen hier nach Stichwörtern geordnet vor. Die in Kapitel 3 beschriebenen Ziele wurden erreicht:

Vertrauensmodell für Agenten: Wir haben eine Theorie entwickelt, wie Agenten trotz einer geringen Menge selbst beobachteter Daten qualifizierte Modelle über das Verhalten anderer Agenten berechnen können. Diese Theorie beruht auf der Nutzung der Kommunikation mit anderen Agenten und dem Austausch von Beobachtungen. Sie berücksichtigt dabei betrügerische Agenten und absichtlich manipulierte Aussagen. Zusätzlich haben wir ein Konzept von Vertrauen entwickelt, das sich auf den Ansatz von Castelfranchi und Falcone stützt.

Spieltheorie und Anwendung: Wir haben eine Experimentalumgebung für heterogene Agentengesellschaften entworfen um diese Theorie zu bestätigen oder zu falsifizieren[1]. Sie baut auf gut untersuchte Modelle der Spieltheorie auf. Dadurch ist

[1] Der Formalismus folgt den Empfehlungen von Karl Popper: Er ist einfach, er grenzt sein Anwendungsgebiet klar ab, die Ergebnisse sind reproduzierbar und überprüfbar und diese

es auch möglich, andere Ansätze mit dieser Arbeit zu vergleichen (z.B. (Axelrod, 1984)). Die Eigenschaften dieser Experimentalumgebung zeigen außerdem deutliche Parallelen zu interessanten Anwendungsszenarien, wie z.B. dem *Electronic Commerce* (und damit auch dem Speditionsszenario), dem Problem der mobilen Agenten und dem *public key management* aus der Kryptographie.

Experimentelle Evaluation: In einer experimentellen Evaluation dieser Theorie haben wir gezeigt, daß Agenten durch Benutzung ihrer Implementierung das Verhalten anderer Agenten wesentlicher schneller und besser approximieren können.

Erweiterbarkeit: Der Formalismus zeigt, daß seine Nutzung nicht auf die Beurteilung des Kooperationswillens anderer Agenten beschränkt ist. Er läßt sich sofort auf alle Eigenschaften eines Agenten erweitern, die sich mit Hilfe einer Dichotomie erfassen lassen, z.B. „Agent verfügt über viele-wenig Ressourcen", „ist ein guter-schlechter Problemlöser", „reagiert schnell-langsam", „ist vorsichtig-unvorsichtig" usw. (siehe auch Abschnitt 5.1.3).

Modellierung von Vertrauen für andere Szenarien: Prof. Castelfranchi bestätigte in einem Gespräch, daß diese Arbeit ein wichtiges Indiz dafür ist, daß die Untersuchung der möglichen Motive für einen Vertrauensbruch eine wichtige Grundlage der Modellierung von Vertrauen ist. Dies zeigt sich in der vorliegenden Arbeit in der Definition des *Betrugs* (siehe Kapitel 4).

7.2. Wissenschaftlicher Beitrag

Soziale kompetentere Agenten: In dieser Arbeit haben wir beschrieben, wie die Kommunikation zwischen Agenten genutzt werden kann, um schneller das Verhalten von anderen Agenten auf Ehrlichkeit und Willen zur Kooperation einschätzen zu können. Die Agenten benötigen dazu weder ein *a priori* Wissen über das Verhalten anderer, noch benötigen sie die Hilfe von absolut benevolenten Agenten. Die Formalisierung ist in Kapitel 4 vorgestellt. Diese Formalisierung ist in der Lage, mit Aussagen betrügerischer Agenten und absichtlich manipulierter Daten umzugehen, in dem sie Redundanz nutzt. Sie ist mathematisch fundiert und kann implementiert werden, um Agenten „sozial kompetenter" zu machen.

Beschleunigung von Lernmechanismen: Dieser Formalismus basiert auf gut untersuchten spieltheoretischen Modellen. Er läßt sich daher sehr gut in andere Arbeiten einfügen. Insbesondere kann diese Arbeit benutzt werden, um das Lernen von Strategien wie es von Carmel & Markovitch oder Mor et al. beschrieben wurde, drastisch zu beschleunigen. Arbeiten wie die von Bazzan et al. oder Armstrong und Durfee können mit dieser Arbeit so erweitert werden, daß deren Annahme über die Öffentlichkeit der Strategien von Agenten nicht mehr notwendig ist.

Maßstab für Agentenstrategien: Es wurde das *Offen Gespielte Gefangenendilemma mit Partnerauswahl* vorgestellt. Dieses dient als Experimentalumgebung, in der Interaktionen von Agenten analysiert werden können, bei denen Ehrlichkeit und Betrug möglich sind. Als Erweiterung zu dem vielzitierten Turnier von Axelrod haben die

---

Arbeit ist keine "Sackgasse": sie bietet Möglichkeiten der Erweiterung und Verfeinerung (Popper, 1960).

Agenten hier die Möglichkeit, sich ihren Interaktionspartner selbst auszuwählen und sie können mit anderen Agenten über ihre Intentionen kommunizieren.

Erweiterung des bestehenden Vertrauensmodells: Das Vertrauensmodell von Castelfranchi und Falcone wurde in zwei Hinsichten erweitert. Erstens wurde Vertrauen in Kommunikation mit einer eindeutigen Semantik definiert und verwendet, um kommunizierte Inhalte zu bewerten. Zweitens wurde das Modell präzisiert, indem angegeben wurde, wie die Variable *Intention* aus diesem Vertrauensmodell berechnet werden kann.

Erfolgreicher Ansatz: Der Formalismus wurde implementiert und evaluiert. Es hat sich gezeigt, daß in realistischen Szenarien der Anstieg der Performanz nach nur zehn Interaktionen signifikant ist. Nach weiteren zehn Interaktionen beträgt der Performanzanstieg mehr als 15 Prozent. Die Modellqualität wird im selben Zeitraum um zwanzig Prozent verbessert. Die Analyse dieser Arbeit ist in Kapitel 6 dargestellt.

Soziologischer Bezug: Wir haben mit dem Modell von Vertrauen einen „sozialen Kitt" geschaffen, über den innerhalb einer Gruppe Informationen zuverlässig kommuniziert und zur Koordination von Verhalten genutzt werden können. Deshalb ist dieses Modell auch von soziologischem Interesse.

7.3. Ausblick

Mit dem hier präsentierten Modell lassen sich unserer Meinung nach eine Reihe von weiteren Annäherungen an realistische Anwendungen testen.

1. Eine Weiterentwicklung der benutzten Variante wäre, sie als *beschränkt rationale* Agenten zu implementieren. Eine beschränkte Ressource könnte, wie im Abschnitt 5.1.7 beschrieben, die zu speichernde Beobachtungsmenge sein. Solche Agenten müßten dann Entscheidungen darüber fällen, welche Daten sie speichern und welche nicht. Bei sehr großen Gesellschaften ist dies eine sehr kritische Entscheidung, da die Datenmenge wahrscheinlich ebenso groß ist. Außerdem könnte die Anzahl der Agenten, über die sie sich etwas merken können, beschränkt werden. Eine dritte Möglichkeit wäre, das Spiel nicht in Protokollform durchzuführen, sondern alle Agenten gleichzeitig Angebote machen zu lassen. Dann würde der Agent, der schneller seine Entscheidung für einen Interaktionspartner trifft, mit höherer Wahrscheinlichkeit den Zuschlag bekommen. Er müßte dann abwägen, ob er mehr Zeit in eine rechenintensive Berechnung steckt oder ein höheres Risiko eingehen will. Es wäre interessant zu Untersuchen, welche Auswirkungen dabei das (nach Luhmann ja gerade *komplexitätsreduzierende*) Konzept Vertrauen hat.
2. Das Turnier von Axelrod, in dem verschieden Strategien für das Gefangenendilemma gegeneinander angetreten sind, hat unter anderem den Nachteil, daß alle Agenten zu den Interaktionen gezwungen werden. Diesen Nachteil behebt das *Offen Gespielte Gefangenendilemma mit Partnerwahl*. Deshalb liegt eine Neuauflage von Axelrods Turnier nahe, diesmal aber mit der Möglichkeit zur Kommunikation und Partnerauswahl. Eine wichtige Rolle in einem solchen Turnier würde ein *tit for tat* spielen, das Vertrauen benutzt. Theoretisch müßte es wesentlich

besser abschneiden, da es sich besser vor den egoistisch spielenden Agenten schützen kann.

3. Im vorliegenden Modell sind die Matrizen über die Zeit konstant. Würden diese variabel gemacht, ließen sich Zeiten hoher und niedriger Erträge modellieren. Damit könnten wir die Veränderung der Vertrauenswürdigkeit in Abhängigkeit von der Höhe der zu gewinnenden Punkte studieren und könnten untersuchen, wie sich Spieler gegen strategisch spielende Agenten schützen könnten. Außerdem könnte das eigene Spielverhalten von der Einschätzung des Gegenübers abhängig gemacht werden. So könnten z.B. Altruisten mit einem Egoisten egoistisch spielen, selbst wenn sie die Wahl haben, ein Spiel mit ihm zu verweigern.
4. Die Agentengesellschaft, die hier untersucht wurde, war zwar *heterogen*, aber die Werte der Konfigurationen waren gleich verteilt. Möglich wäre die Untersuchung von Agentengesellschaften, die verschiedene Verteilungen von Gesellschaften haben (Glockenkurve, umgekehrte Glockenkurve etc.).
5. Die Agenten könnten ihre Zeugen noch wesentlich intelligenter auswählen. Sie könnten z.B. die letzten mitgeteilten Daten analysieren, um die Wahrscheinlichkeit zu bestimmen, mit der ein Zeuge überhaupt etwas (Neues) weiß. Außerdem könnte eingeschränkt werden, über wieviel Agenten in einer Anfrage Daten angefordert werden. Die Agenten sollten weiterhin in der Lage sein, Zeugen aufgrund ihrer Aussagen statt nur ihres Spielverhaltens zu bewerten. Interessant wäre auch, zu untersuchen, wie Agenten reagieren sollen, wenn sie versuchen, mit Agenten zu kommunizieren, die sich weigern zu antworten
6. Schließlich wäre es wichtig zu untersuchen, welche Arten von Betrug es bei Zeugenaussagen noch geben kann und wie ein „Kreuzverhör" von mehreren Zeugen dazu dienen kann, diese aufzudecken oder abzuschätzen.

Interessant wäre auch, Simulationen mit größeren Gesellschaften und der Möglichkeit der Koalitionsbildung durchzuführen. Gerade mit der Untersuchung der Koalitionsbildung ergeben sich eine Reihe spannender, neuer Fragen für die Soziologie und die Informatik.

Abkürzungen

AOP	Agent Oriented Programming
CMI	Computer Mediated Interaction
CNP	Contract Net Protocol
COMRIS	CO-habited Mixed-Reality Information Spaces
CORBA	Object Request Broker
CSCW	Computer Supported Collaborative Work
ESS	Evolutionary Stable Strategies
FIPA	Foundation for Intelligent Physical Agents
HCI	Human Computer Interaction
KI	Künstliche Intelligenz
KNN	Künstliche Neuronale Netze
KQML	Knowledge Query and Manipulation Language
MAS	Multi-Agenten Systeme
OGGD	Offen Gespieltes Gefangenendilemma
OISS	Open Information System Semantics
OMG	Object Management Group
SIF	Social Interaction Framework
VKI	Verteilte Künstliche Intelligenz

Referenzen

(Agha und Hewitt, 1987) G. A. Agha und C. Hewitt. Concurrent programming using actors. In Yonezawa, A. und Tokoro, M. (Hrsg.), *Object-Oriented Concurrent Programming*. MIT Press, Cambridge, MA, USA, 1987.

(Agre und Chapman, 1987) P. E. Agre und D. Chapman. Pegi: An Implementation of a Theory of Activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*. 1987.

(Armstrong und Durfee, 1998) A. Armstrong und E. H. Durfee. *Mixing and Memory: Emergent Cooperation in an Information Marketplace*. In (Demazeau, 1998).

(Demazeau, 1998) Y. Demazeau. *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS 98)*, 1998.

(Axelrod, 1984) R. Axelrod. *The Evolution of Cooperation*. New York: Basic Books, 1984

(Aylett, 1997) R. Aylett. *Robots as Socially Intelligent Agents*. In (Dautenhahn et al., 1997)

(Bachmann, 1998) R. Bachmann. Kooperation, Vertrauen und Macht in Systemen Verteilter Künstlicher Intelligenz. In *Beiträge für den ersten Workshop "Sozionik"*, Hamburg, 1998.

(Barber, 1983) B. Barber. *Logic and Limits of Trust*. New Jersey: Rutgers Univeristy Press, 1983.

(Baurmann und Mans, 1984) M. Baurmann und D. Mans. Künstliche Intelligenz in den Sozialwissenschaften: Expertensysteme als Instrumente der Einstellungsforschung. In *Analyse und Kritik*, Opladen: Westdeutscher Verlag, S. 103-159, 1984.

(Bazzan et al., 1997) A. L. C. Bazzan, R. H. Bordini und J. A. Campbell. *Agents with Moral Sentiments in an Iterated Prisoner's Dilemma Exercise*. In (Dautenhahn et al., 1997).

(Bertino et al., 1996) E. Bertino, H. Kurth, G. Martella und E. Montolivo. *Computer Security- ESORICS96; 4th European Symposium on Reasearch in Computer Securiy*. Rome, Italy; Lecture Notes in Computer Science, Springer, 1996.

(Beth et al., 1994) T. Beth, M. Borcherding und B. Klein. Valuation of Trust in Open Networks. In *Computer Security-ESORICS 94; 3rd European Symposium on research in Computer Security*. Brighton, UK, November, 1994.

(Bibel, 1995) W. Bibel. Identität und Vision der Künstlichen Intelligenz. BMBF (Hrsg.). *Mit leisen Schritten- von der Künstlichen Intelligenz als Vision zur intelligenten Technik als Perspektive*, Veranstaltung des Bundesministeriums für Forschung und Technologie in Zusammenarbeit mit dem VDI/VDE-Technologiezentrum Informationstechnik Teltow, Bonn, S. 18-21, 1995.

(Biswas et al., 1999a) A. Biswas, M. Mundhe, S. Debnath und S. Sen. *A Bayesian Network based Approach for Modeling Agent Relationships*. In (Castelfranchi et al., 1999a).

(Biswas et al., 1999b) A. Biswas, S. Sen und S. Debnath. *Limiting Deception in Social Agent-Groups*. In (Castelfranchi et al., 1999a).

(Blumenberg, 1994) B. Blumenberg. Action-selection in Hamsterdam: Lessons from ethology. In *Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior.*, 1994.

(Bond und Gasser, 1988) A. Bond und L. Gasser. *Readings in Distributed Artificial Intelligence.* Morgan Kaufmann, Los Angeles, CA, 1988.

(Bratman et al., 1987) M. E. Bratman, D. J. Israel und M. E. Pollack. *Towards an Architecture for Resource-bounded Agents.* Technical Report CSLI-87-104, Center for the Study of Language and Information, SRI and Stanford University, August, 1987.

(Brainov und Sandholm 1999) S. Brainov und T. Sandholm. *Contracting with uncertain level of trust*. In (Castelfranchi et al., 1999a).

(Bronstein und Semendjajew 1991) I. N. Bronstein, K. A. Semendjajew. *Taschenbuch der Mathematik*. B. G. Teubner Verlagsgesellschaft, 1991.

(Brooks, 1986) R. Brooks. A robust layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation* RA-2, S. 14-23, 1986.

(Brooks, 1991) R. Brooks. Intelligence without Representation. In *Artificial Intelligence* **47**, S. 139-159, 1991.

(Boon und Holmes, 1991) S. Boon und J. G. Holmes. The dynamics of interpersonal trust: resolving uncertainty in the face of risk. In A. Hinde und J. Groebel (Hrsg.), *Cooperation and Prosocial Behaviour.* S. 190-211, Cambridge University Press, 1991.

(Brückerhoff, 1982) A. Brückerhoff. *Vertrauen. Ein Versuch einer phänomenologisch-idiographischen Näherung an ein Konstrukt.* Dissertation, Universität Münster, 1982.

(Bürckert et al., 2000) Hans-Jürgen Bürckert, Gero Vierke und Petra Funk. An Intercompany Dispatch Support System for Intermodal Transport Chains. In *Intelligent Systems in Traffic and Transportation: Decision Technologies for Management-Track of the Thirty-Third Hawaii*

International Conference on System Sciences (HICSS-33), Hawaii, Januar 2000, in Vorbereitung.

(Carmel und Markovitch, 1998) D. Carmel und S. Markovitch. *How to Explore Your Opponent's Strategy (almost) Optimally*. In (Demazeau, 1998).

(Castelfranchi et al., 1997) C. Castelfranchi, F. de Rosis und R. Falcone. *Social Attitudes and Personalities in Agents*. In (Dautenhahn et al., 1997)

(Castelfranchi und Falcone, 1998) C. Castelfranchi und R. Falcone. *Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification*. In (Demazeau, 1998).

(Castelfranchi und Falcone, 1999) C. Castelfranchi und R. Falcone. *The Dynamics of Trust: From Beliefs to Action*. In (Castelfranchi et al., 1999a).

(Castelfranchi et al., 1999a) C. Castelfranchi, Y. Tan, R. Falcone, B. S. Firozabadi. *Deception, Fraud and Trust in Agent Societies. Proceedings of the Workshop at the Autonomous Agents Conference 1999*. National Research Council, Institute of Psychology, Rome, Italy, 1999.

(Chaib-Draa, 1994) B. Chaib-Draa. Distributed Artificial Intelligence: An Overview. In *Encyclopedia of Computer Science and Technology*, Bd. 31, S. 215-242, Marcel Dekker, Inc., 1994.

(Charniak, 1991) E. Charniak. Bayesian Networks without Tears. In *AI Magazine*. S. 50-63, 1991.

(Chinneck, 1999) J. W. Chinneck. *How to Organize your Thesis*. Verfügbar per http://www.sce.caleton.ca/faculty/chinneck/thesis.html.

(Christianson und Harbison, 1997) B. Christanson und W. S. Harbison. Why isn't Trust Transitive? In *Security Protocols,* Lecture Notes in Computer Science, 1189, Springer-Verlag, Berlin, 1997.

(Cohen und Levesque, 1985) P. R. Cohen und H. J. Levesque *Speech Acts and rationality. 23rd Annual Meeting*, 1985.

(Cohen und Levesque, 1987) P. R. Cohen und H. J. Levesque. Intention = Choice + Commitment. In *Proceedings AAAI-87*, S. 410-415, Seattle, WA, 1987.

(Cohen und Levesque, 1990) P. R. Cohen und H. J. Levesque. Intention is Choice with Commitment. *Artificial Intelligence* **42**, S. 213-261, 1990.

(Coleman, 1990) J. Coleman. *The Foundations of Social Theory*. The Belknap Press of the University of Harvard., 1990.

(Conamero und Van de Velde, 1997) D. Conamero und W. Van de Velde. *Socially Emotional: Using Emotions to Ground Social Interaction*. In (Dautenhahn et al., 1997).

(Collins, 1992) H. M. Collins. Forms of Life and a Simple Test for Machine Intelligence. In *Social Studies of Science*, **22**, S. 726-739, 1992.

(Dasgupta, 1990) P. Dasgupta. Trust as a Commidity. In (Gambetta, 1990c) S. 49-72, Blackwell, 1990.

(Dautenhahn et al., 1997) K. Dautenhahn, J. Masthoff und C. Numaoka. *Socially Intelligent Agents*. Papers from the, 1997 AAAI Fall Symposium, November 8-10, Cambridge, Massachusetts, Technical Report FS-97-02, 1997.

(Daws und Thaler, 1988) R. Daws und R. Thaler. Anomalies: Cooperation. *Journal of Economic Perspectives* **2(3)**, 1988.

(Demiris und Hayes, 1997) J. Demiris und G. Hayes. *Do Robots Ape?* In (Dautenhahn et al., 1997).

(Doran, 1998) J. Doran. *Social Simulation, Agents and Artificial Societies.* In (Demazeau, 1998).

(Delahaye und Mathieu, 1998) J. P. Delahaye und P. Mathieu. Altruismus mit Kündigungsmöglichkeit. In *Spektrum der Wissenschaft*, ISSN 0170-2971, Februar, 1998.

(Deutsch, 1960) M. Deutsch. Trust, Trustworhiness, and the F-Scale. In *Jounral of Abnormal and Social Psychology* **61**, 1960.

(Deutsch, 1962) M. Deutsch. Cooperation and Trust: Some Theoretical Notes. In M. R. Jones (Hrsg.), *Nebraska Symposium on Motivation*. Nebraska University Press, 1962.

(Deutsch, 1973) M. Deutsch. *The Resolution of Conflict*. New Haven and London: Yale University Press, 1973.

(Durfee, 1991) E. Durfee, (Hrsg.), Special issue on DAI:, 10 years later of the *IEEE Transactions on Systems, Man and Cybernetics*. **21**, November / Dezember, 1991.

(Durfee und Rosenschein, 1994) E. Durfee und J. Rosenschein. Distributed Problem Solving and Multi-Agent Systems: Comparison and Examples. *Proceedings of the, 13th International Workshop on Distributed Artificial Intelligence*, S. 94-104, 1994.

(Edelmann, 1987) G. Edelmann. *Neural Darwinism: The Theory of neural Group Selection*. Basic Books, 1987.

(Finin et al. 93) T. Finin, J. Weber, G. Wiederhold, M. Genesereth, R. Fritzson, J. McGuive, S. Shapiro und C. Beck. *Specification of the KQML Agent Communication Language.*, DARPA Knowledge Sharing Initiative: External Interfaces Group, University of Maryland, Juni, 1993.

(Fischer et al., 1998) K. Fischer, C. Ruß und G. Vierke. *Decision Theory and Coordination in Multi-Agent Systems*. Research Report RR-98-02, DFKI Saarbrücken, 1998.

(Florian, 1996) M. Florian. "Soziomedia" als Virtualisierung von Kultur? Überlegungen zur Sozialmetaphorik in der KI an der Grenze zwischen Realität und Virtualität. In B. Becker, C. Lischka und J. Wehner (Hrsg.): *Kultur - Medien-Künstliche Intelligenz. Beiträge zum Workshop während der, 19. Jahrestagung für Künstliche Intelligenz*. GMD-Studien Nr. 290, Mai, 1996.

(Freeman und Skapura, 1991) J. A. Freeman und D. M. Skapura. *Neural Networks*. Addison Wesley, 1991.

(Fudenberg und Tirole 1991) D. Fudenberg und J. Tirole. *Game Theory*. MIT Press, 1991.

(Funk und Lind, 1997) P. Funk und J. Lind. *What is a Friendly Agent?* In (Dautenhahn et al., 1997)

(Gambetta 1990a) D. Gambetta (Hrsg.). *Trust*. Blackwell,1990.

(Gambetta, 1990b) D. Gambetta. *Can we trust Trust?* In (Gambetta 1990a).

(Gambetta, 1990c) D. Gambetta. *Mafia: The Price of Distrust*. In (Gambetta 1990a), Seiten 158-176.

(Gasser, 1991) L. Gasser. Social Conceptions of Knowledge and Action: DAI Foundations and Open System Semantics. In *Artificial Intelligence* **47**, 1991.

(Georgeff und Ingrand, 1990) M. P. Georgeff und F. F. Ingrand. Decision-making in Embedded Reasoning Systems. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, S. 972-978, 1990.

(Genesereth et al., 1984) M. R. Genesereth, M. L. Ginsberg, and J. Rosenschein. *Cooperation without Communication*. Technical Report 84-36, Stanford Heuristic Programming Project, Computer Science Department, Stanford University, 1984.

(Golembiewski und McConkie, 1975) R. T. Golembiewski und M. McConkie. The Centrality of Interpersonal Trust in Group Processes. Cary L. Cooper (Hrsg.): *Theories of Group Processes*. S. 131-185, Wiley, 1975.

(Good, 1971) I. Good. Twenty-seven Principles of Rationality. In V. Godambe und D. Sprott (Hrsg.): *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, Winston, 1971.

(Green et al., 1997) S. Green, L. Hurst, B. Nagle, P. Cunningham, F. Sommers und R. Evans. *Software agents: A review*. IAG report, Trinity College

Dublin, Broadcom Éreann Research, Intelligent Agents Group, 1997.

(Halpin, 1998) B. Halpin. Computer Simulation in Sociology: A Review. In *American Behavioural Scientist*, 1998.

(Henecka, 1994) H. P. Henecka. *Grundkurs Soziologie*. Leske + Budrich, Opladen, 1994.

(Simon, 1955) I. Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* **59**, S. 99-118, 1955.

(Hertzberg, 1989) J. Hertzberg. *Planen: Einführung in die Planerstellungsmethoden der künstlichen Intelligenz.B.*I. Wissenschaftsverlag, 1989.

(Hewitt, 1991) C. Hewitt. Open Information Systems Semantics for Distributed Artificial Intelligence. *Artificial Intelligence* **47**, S. 79-116, 1991.

(Hewitt und de Jong, 1984) C. Hewitt und P. de Jong. Open Systems. In M. Brodie (Hrsg.). *On conceptual modelling*. Springer, New York, 1984.

(Jones 90) S. Jones. *A Discussion of Issues and Systems Relevant to Computer Supported Cooperative Work*. Technical Report 64. University of Stirling, Department of Computing Science and Mathematics, 1990.

(Kahan und Rapoport, 1984) J. P. Kahan und A. Rapoport. *Theories of Coalition Formation*. Lawrence Erlbaum Associates Publishers, 1984.

(Ketchpel, 1994) S. Ketchpel. Forming Coalitions in the Face of Uncertain Rewards. In *AAAI*, S. 414-419, Seattle, Washington, Juli, 1994.

(Koller, 1990) M. Koller. *Sozialpsychologie des Vertrauens: Ein Überblick über theoretische Ansätze*. Bielefelder Arbeiten zur Sozialpsychologie, 1990.

(Lind, 1998) J. Lind. *The EMS model*. DFKI Technical Memo TM-98-09, Saarbrücken, 1998.

(Luce und Raiffa, 1957) R. D. Luce und H. Raiffa. *Games and Decisions, Introduction and Critical Survey*. Wiley, New York, 1957.

(Luger, 1994) G. F. Luger. *Cognitive Science: The Science of Intelligent Systems*. Academic Press, 1994.

(Luhmann, 1973) N. Luhmann. *Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart, 1973.

(Luhmann, 1979) N. Luhmann. *Trust and Power*. Chichester, Wiley, 1979.

(Luhmann, 1990) N. Luhmann. *Familiarity, Confidence, Trust: Problems and Alternatives*. In (Gambetta 1990a), S. 94-107.

(Lux und Steiner, 1995) A. Lux und D. Steiner. Understanding cooperation: an agent's perspective. In *Proceedings of the First International, Conference on Multi-Agent Systems (ICMAS '95), 1*995.

(Maes, 1989) P. Maes. *How to do the right thing.* AI Memo 1180, MIT AI Laboratory, 1989.

(Malsch et al., 1996) T. Malsch, M. Florian, M. Jonas und I. Schulz-Schäfer. *Expeditionen ins Grenzgebiet zwischen Soziologie und Künstlicher Intelligenz.* In Künstliche Intelligenz **2**, S. 6-12, 1996.

(Marsh, 1994) S. P. Marsh. *Formalising Trust as a Computational Concept.* Phd Thesis, Department of Computing Science and Mathematics, University of Stirling, 1994.

(von Martial, 1992) F. v. Martial. *Einführung in die Verteilte KI.* In Künstliche Intelligenz **1**, S. 6-11, 1992.

(Maurer, 1996) U. Maurer. *Modelling a Public-Key Infrastructure.* In (Bertino et al., 1996).

(Minsky, 1961) M. Minsky. Steps towards artificial intelligence. *Proceedings of the IRE,* Seite 8-30, 1961. Neu erschienen in E.A. Feigenbaum und Feldman, J. (Hrsg.), *Computers and Thought*, S. 406-450, McGraw-Hill, 1963.

(Minsky 1986) M. Minsky. *The society of mind.* Simon and Schuster, New York, 1986.

(Mitchell 1997) T. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

(Morik, 1989) K. Morik. Sloppy Modelling. In *Knowledge Representation and Organization in Machine Learning.* Springer Verlag, 1989.

(Mor et al., 1996) Y. Mor, C. V. Goldman und J. S. Rosenschein. *Learn Your Opponent's Strategy (in Polynomial Time)!* In (Weiß und Sen, 1996).

(Muir, 1987) B. Muir. Trust between Humans and Machines, and Design of decision Systems. In *International Journal of Man-Machine Studies* **27**, Seiten 527-539, 1987.

(Müller, 1993) J. Müller. *Verteilte Künstliche Intelligenz: Methoden und Anwendungen.* BI-Wissenschaftsverlag, 1993

(Müller, 1996) J. Müller. *The Design of Intelligent Agents: A Layered Approach*, Lecture Notes in Artificial Intelligence 1177. Springer Verlag, 1996.

(Müller und Pischel, 1993) J. Müller und M. Pischel. InteRRaP: eine Architektur zur Modellierung Flexibler Agenten. In H. J. Müller, (Hrsg.), *Beiträge zum Gründungsworkshop der Fachgruppe VKI.* DFKI Saarbrücken, April, 1993.

(Nash 1950) J. Nash, The bargaining problem. *Econometrica* 18, S. 155-162, 1950.

(Narowski, 1974) C. Narowski. *Vertrauen. Begriffsanalyse und Operationalisierungsversuch*. Dissertation, Universität Tübingen, 1974.

(Osborne und Rubinstein, 1994) M. Osborne und A. Rubinstein. *A course in Game Theory*. MIT Press, 1994.

(Pearl 1988) J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, 1988.

(Pearl, 1997) J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, California: Morgan Kaufmann, 1997.

(Platzköster, 1990) M. Platzköster. *Vertrauen: Theorie und Analyse interpersoneller, politischer und betrieblicher Implikationen*. Verlag Beleke KG, 1990.

(Pollack und Ringuette, 1990) M. Pollack und M. Ringuette. Introducing the TILE-WORLD: Experimentally evaluationg agent architectures. In *Proceedings of the Conference of the American Association for Artificial Intelligence*, S. 183-189, 1990.

(Popper, 1960) K. Popper. *The Logic of Scientific Discovery*. Hutchinson of London, 1960.

(Preece, 1994) J. Preece. *Human-Computer Interaction*. Addison-Wesley, 1994.

(Rao und Georgeff, 1991) A. S. Rao und M. P. Georgeff. Modeling Agents Within a BDI-Architecture. In R. Fikes und E. Sandewall (Hrsg.): *Proceedings of the 2nd Interantionl Conference on Principles of Knowledge representation and Reasoning (KR'91)*, S. 473-484, Cambridge, Mass., Morgan Kaufmann, April, 1991.

(Rao et al., 1992) A. S. Rao und M. P. Georgeff und E. A. Sonenberg. Social Plans: A Preliminary Report. In Y. Demazeau und E. Werner (Hsg.), *Decentralized AI* 3, S. 57-76, 1992.

(Rapoport und Orwant, 1962) A. Rapoport und C. Orwant. Experimental games: a Review. *Behavioural Sciences* **7**, 1962.

(Rapoport und Chammah, 1970) A. Rapoport und Albert M. Chammah. *Prisoner's Dilemma; A Study in Conflict and Cooperation*. University of Michigan Press, Ann Arbor, 1970.

(Rempel und Holmes, 1986) J. Rempel und J. Holmes. How do I Trust Thee? *Psychology Today*, Februar, S. 28-34., 1986.

(Rosenschein, 1985) J. S. Rosenschein. *Rational Interaction: Cooperation among Intelligent Agents*. Ph.D. Thesis, Stanford University, 1985.

(Rosenschein und Genesereth, 1985) J. S. Rosenschein und M. R. Genesereth. Deals among rational agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, S. 91-99, 1985.

(Rosenschein und Kaebling, 1986) J. S. Rosenschein und L. P. Kaebling. The synthesis of digital machines with provable epistemic properties. In *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*. Morgan Kaufmann, 1986.

(Rosenschein und Zlotkin, 1994) J. S. Rosenschein und Gilad Zlotkin. *Rules of Encounter*. Cambridge, Mass., MIT Press, 1994.

(Rotter, 1971) J. B. Rotter. Generalized Expectancies for Interpersonal Trust. In *American Psychologist* **26**, S. 443-452, 1971.

(Ruß, 1997) C. Ruß. *Economic Mechanism Design for the Auchtion-Based Coordination of Self-Interested Agents*. Diplomarbeit, Fachbereich Informatik, Universität des Saarlandes, 1997.

(Russell und Norvig, 1996) S. Russel und P. Norvig. *Artificial Intelligence, A Modern Approach*. Prentice-Hall, 1996.

(Russell und Wefald, 1991) S. Russell und E. Wefald. *Do the Right Thing*. MIT Press, 1991.

(Sandholm, 1998) T. Sandholm. *Agents in Electronic Commerce: Component technologies for Automated Negotiation and Coalition Formation*. In (Demazeau, 1998).

(Sandholm und Lesser, 1995) T. Sandholm und V. Lesser. Coalition formation among bounded rational agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

(Searle, 1969) J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

(Seligmann, 1997) A. Seligman. *The problem of Trust*. Princeton University Press, 1997.

(Sengers, 1996) P. Sengers. Socially situated AI: What it is and why it matters. In H. Kitano, (Hrsg.): *AAAI-96 Workshop on AI/ A-Life and Entertainment*. Menlo Park, CA:AAAI Press. AAAI Technical Report WS-96-03, 1996.

(Schillo et al., 1999a) M. Schillo, J. Lind, P. Funk, C. Gerber und C. Jung. *SIF - The Social Interaction Framework. System Description and User's Guide to a Multi-Agent System Testbed.* DFKI Research Report RR-99-02, Saarbrücken, 1999.

(Schillo et al., 1999b) M. Schillo, P. Funk und M. Rovatsos. Who can you Trust: Dealing with Deception. In Rino Falcone (Hrsg.) *Proceedings of the Workshop "Deception, Fraud and Trust" of the Autonomous Agents Conference*, 1999.

(Schillo und Funk, 1998) M. Schillo und P. Funk. Spontane Gruppenbildung in künstlichen Gesellschaften. In *Proceedings des Workshops Sozionik der 22. Jahrestagung fuer Kuenstliche Intelligenz*, 1998.

(Schillo und Funk, 1999) M. Schillo und P. Funk. Learning form and about other Agents in Terms of Social Metaphors. In J. M. Vidal and S. Sen (Hrsg.) *Proceedings of the "Agents Learning About, From and With other Agents" Workshop of the 16th International Joint Conference on Artificial Intelligence (IJCAI)* , 1999.

(Schum, 1981) D. A. Schum. Sorting out the Effects of Witness Sensitivity and Response-Criterion Placement upon the Inferential Value of Testimonial Evidence. In *Organizational behavior and human performance* **27**, S. 153-196, 1981.

(Shafer und Pearl, 1990) G. Shafer und J. Pearl. *Readings in uncertain reasoning*. The Morgan Kaufmann series in Representation and Reasoning, Morgan Kaufmann, San Mateo, CA, 1990.

(Shechory und Kraus, 1993) O. Shechory und S. Kraus. Coalition Formation among Autonomous Agents: Strategies and Complexity. In C. Castelfranchi und J. P. Müller (Hrsg.): *Proceedings of the fifth European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW 93)*, Lecture Notes in Computer Science 957, Springer, 1993.

(Shoham, 1993) Y. Shoham. Agent-Oriented Programming. In *Artificial Intelligence* **60**, S. 51-92, 1993.

(Smith, 1982) J. Maynard Smith. Evolution and the Theory of Games. Cambridge University Press, Cambridge, UK, 1982.

(Smith, 1980) R. G. Smith. The Contract-Net Protocol: High-Level Communication and Control in a Distributed Problem Solver. *IEEE Trans. Computers*, S. 1104-1113, 1980.

(Straffin 1996) P. D. Straffin, *Game Theory and Strategy*, MIT Press, 1996.

(Suchman, 1987) L. A. Suchman. *Plans and Situated Actions*. Cambridge University Press, Cambridge, 1987.

(Sundermeyer, 1993) K. Sundermeyer. Modellierung von Agentensystemen. In J. Müller *Verteilte Künstliche Intelligenz: Methoden und Anwendungen*. BI-Wissenschaftsverlag, 1993

(Tan und Thoen 1999) Y. Tan und W. Thoen. *Towards a Generic Model of Trust for Electronic Commerce*. In (Castelfranchi et al., 1999a)

(Turing, 1950) A. M. Turing. Computing Machinery and Intelligence. Mind LIX no. 2236, S. 433-60, Oktober, 1950. (auch in M. A. Boden (Hrsg.): *The Philosophy of Artifical Intelligence*, Oxford University Press, 1990.)

(van den Berghe, 1980) P. van den Berghe. The Human Family: A Sociobiological Look. In Joan S. Lockard (Hrsg.): *The Evolution of human Social Behaviour*. Kapitel 4, S. 67-85, New York, Elsevier, 1980.

(van der Linden und Verbeek, 1985) W. J. van der Linden und Albert Verbeek. Coalition formation: a gametheoretic approach. In Henk A. M. Wilke (Hrsg.): *Coalition Formation,* Advances in Psychology **24**. North Holland, 1985.

(Vigna, 1998) G. Vigna. *Mobile Agents and Security*. Lecture Notes in Computer Science 1419, Springer Verlag, 1998.

(von Neumann und Morgenstern, 1944) J. von Neumann und O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1944.

(Vossen, 1997) G. Vossen. The CORB*A* Specification for Cooperation in Heterogeneous Information Systems. In Peter Kandyia und Matthias Klusch (Hrsg.): *Proceedings of the First International Workshop on Cooperative Information Agents*. S. 101-115, Springer Verlag, 1997.

(Watzlawick, 1997) P. Watzlawick. *Wie wirklich ist die Wirklichkeit?* Serie Piper, München, 1997.

(Weiß, 1996) G. Weiß, *Adaptation and Learning in multi-agent systems: Some Remarks and a Bibliography*. In (Weiß und Sen, 1996).

(Weiß und Sen, 1996) G. Weiß und S. Sen. *Adaptation and Learning in Multi-Agent Systems*: *Proceedings of the IJCAI'95 Workshop*. Lecture Notes in Artificial Intelligence, Springer, 1996.

(Wille, 1993) R. Wille. Conceptual Lattices and Conceptual Knowledge Systems. *Computers & Mathematics with Applications*, 23:493-515, 1993.

(Wooldridge, 1995) M. Wooldridge. This is MyWorld: The Logic of an Agent-Oriented DAI Testbed. In M. Wooldridge und N. Jennings (Hrsg.): *Intelligent Agents: Proceedings of the 1994 Workshop on Agent Thoeries, Architectures, and Languages*. Lecture Notes in Artificial Intelligence. Springer Verlag, 1995.

(Wooldridge und Jennings, 1995) M. Woolridge und N. Jennings. Intelligent Agents: Theory and Practice. In *Knowledge Engineering Review* 10:2, Cambridge University Press, 1995.

(Xiang, 1994) Y. Xiang. Distributed multi-agent probabilistic Reasoning with Bayesian Networks. In Zbigniew W. Ras and Maria Zemankova (Hrsg.): *Proceedings of the 8th International Symposium on Methodologies for Intelligent Systems*. Charlotte, NC, USA (1994). Lecture Notes in Artificial Intelligence 869. Springer Verlag, Berlin, 1994.

(Young, 1975) O. R. Young. *Bargaining: formal theories of negotiation*. University of Illinois Press., 1975.

(Zacharia, 1999) G. Zacharia. *Trust Management through Reputation Mechanisms.* In (Castelfranchi et al., 1999a).

(Zand, 1977) D. E. Zand. Vertrauen und Problemlösungsverhalten von Managern. In H. E. Lück (Hrsg.): *Mitleid-Vertrauen-Verantwortung*, Stuttgart, 1977.

(Zeng und Sycara, 1996) D. Zeng und K. Sycara. Bayesian Learning in Negotiation. In *Proceedings of the AAAI Stanford Spring Symposium Series on Adaptation, Co-evolution and Learning in Multiagent Systems*, 1996.

(Zimmermann, 1995) P. Zimmermann. *The Official PGP User's Guide*, MIT Press, 1995.

(Zlotkin und Rosenschein, 1994) G. Zlotkin und Jeffrey S. Rosenschein. Coalition cryptography and stability: Mechanisms for coalition formation in task oriented domains. In *AAAI*, S. 432-437, Seattle, Washington, Juli, 1994

Index

Vertrauen und Betrug in Multi-Agenten Systemen

Erweiterung des Vertrauensmodells von Castelfranchi und Falcone um eine Kommunikationskomponente

Michael Schillo

RR-00-02

Research Report