Albert-Ludwigs-Universität Freiburg im Breisgau
Institut für Informatik und Gesellschaft
Abteilung Telematik, Lehrstuhl Prof. Dr. Günter Müller

Studienarbeit

Trust in Bots: Zur Rolle von Vertrauen und Reputation bei Multiagenten-Systemen

Till Westermayer
Habsburgerstr. 44, 79104 Freiburg
till@tillwe.de

Soziologie (HF), Informatik und Psychologie (NF), 10. Semester

19.8.2000

Abstract

Vertrauen ist nicht nur ein gesellschaftliches Phänomen, sondern wird im Bereich der Multiagenten-Systeme (MAS) auch als Mittel zur Lösung sozio-technischer Sicherheitsprobleme diskutiert. In dieser Arbeit werden verschiedene soziologische Ansätze zur Diskussion um Vertrauen und Reputation dargestellt. Die Übertragbarkeit des gesellschaftlichen Vertrauensmodells auf technische Systeme wird diskutiert und Kriterien genannt, wann es auf MAS übertragbar sein kann. Verschiedene Formalisierungen von Vertrauen werden vorgestellt, um schließlich für das am IIG Freiburg entwickelte MAS AVALANCHE Hinweise zur Implementierung eines Reputationsmodells zu geben.

Inhaltsübersicht

1 Einleitung 3

2 Das soziologische Vorbild 5

2.1 Das Vertrauen der Gesellschaft 6

2.2 Gesellschaftlicher Zusammenhalt durch Vertrauen 8

2.3 Vertrauen und kooperatives Handeln 12

2.4 Zusammenfassung 15

3 Zum Zusammenhang von Theorie und Praxis 15

3.1 Ist Vertrauen auf Agentensysteme anwendbar? 15

3.2 Zwischen Simulation und technischer Implementierung 18

3.3 Schlussfolgerungen in Bezug auf AVALANCHE 20

4 Formalisierungen von Reputation und Vertrauen 22

4.1 Agentenzentrierte, solitäre Ansätze 25

4.1.1 Stephen Marsh – Trust mit großem T 26

4.1.2 Weitere agentenzentrierte, solitäre Ansätze 33

4.1.3 Zusammenfassung agentenzentrierte, solitäre Ansätze 35

4.2 Agentenzentrierte, soziale Ansätze 35

4.2.1 Michael Schillo – *TrustNet* oder: Du bist nicht alleine 35

4.2.2 Weitere agentenzentrierte, soziale Ansätze 41

4.2.3 Zusammenfassung agentenzentrierte, soziale Ansätze 45

4.3 ‘Objektive’ externe Bewertungsagenturen 45

4.4 ‘Subjektive’ externe Bewertungsagenturen 47

4.5 Andere Ansätze 51

4.6 Zusammenfassung 53

5 Hinweise zur Implementierung in AVALANCHE 53

5.1 Beschreibung des in AVALANCHE verwendeten Modells 53

5.2 Diskussion und Empfehlungen 55

5.3 Zusammenfassung 59

6 Literatur 60

Trust in Bots: Zur Rolle von Vertrauen und Reputation bei Multiagenten-Systemen

> If you were going to be successful in the criminal world, you needed a reputation for honesty. Terry Pratchett, *Feet of Clay*

1 Einleitung

Multiagenten-Systeme

Seit einigen Jahren ist in der Informatik ein neuerlicher Paradigmenwechsel zu beobachten: verteilte Anwendungen – etwa auf dem Internet – werden nicht mehr in der Begrifflichkeit sich gegenseitig aufrufender Programmteile und Algorithmen beschrieben, sondern mit der Metapher des autonomen Agenten, der mit anderen Agenten – zu denen in manchen Fällen auch die NutzerInnen zählen – interagiert, also Nachrichten austauscht, Handel betreibt oder sogar Nachkommen zeugt. Werden diese 'lebendigen' Agenten im Plural betrachtet, ist die Rede von einem *Multiagenten-System* (MAS).

Informatik ⇔Soziologie

Multiagenten-Systeme haben nicht nur den Reiz des Neuen, und bieten gerade im Anwendungsfeld *eCommerce* erhebliche Vorteile, sondern verfügen auch über gewisse Ähnlichkeiten mit menschlichen Gesellschaften, zumindest insofern, als es hier wie dort eine Ebene einzelner, mehr oder weniger zielgerichteter Handlungen und eine Ebene emergenter, gesellschaftlicher Phänomene gibt. Dementsprechend findet zur Zeit, insbesondere unter dem Label 'Sozionik', ein gegenseitiger Informationsaustausch zwischen (Teilen) der Soziologie einerseits und (Teilen) der Informatik andererseits statt (Malsch 1997; Malsch et al. 1998; vgl. auch Manhart 1999). Aus Sicht der Soziologie geht es dabei darum, herauszufinden, wieweit Modellierungen und Simulationen von Gesellschaften im Computer zu Erkenntnisgewinn über menschliche Gesellschaften beitragen können. Aus informatischer Sicht wird in der Soziologie nach Lösungen für Probleme künstlicher Gesellschaften gesucht, die menschlichen Gesellschaften anscheinend gefunden haben.

Sicher durch Vertrauen?

Eines dieser Problemfelder ist die Frage der Sicherheit eines Multiagenten-Systems, bezogen darauf, wie unkooperatives und betrügerisches Verhalten vermieden werden kann. Eine in der Soziologie gefundene Lösungsmöglichkeit für dieses Problem in menschlichen Gesellschaften scheint *Vertrauen* zu sein. Es stellt sich also u.a. die Frage, wie Vertrauen in menschlichen Gesellschaften erzeugt wird, welche Faktoren an der Bildung einer Reputation beteiligt sind, und wie Hinweise auf Vertrauenswürdigkeit verarbeitet werden. In gewisser Weise könnte selbst bei einem geschlossenen Multiagenten-System, an dem nur gutwillige Agenten beteiligt sind, von der Existenz impliziten Vertrauens zwischen den beteiligten Agenten und in Bezug auf das System gesprochen werden. Anders sieht es aus, wenn ein Multiagenten-System offen ist, oder wenn das Verhalten weniger gutwilliger Agenten simuliert werden soll, wenn also über die Absichten eines unbekannten Agenten prinzipiell nichts bekannt ist. Soll mit diesen interagiert werden, z.B. Handel getrieben

werden, oder eher nicht? Ab wann ist ein Agent vertrauenswürdig genug, um ihn bei der Auswahl von Interaktionspartnern zu berücksichtigen? Wie lässt sich Vertrauenswürdigkeit operationalisieren? Und: wird ein Multiagenten-System vertrauenswürdig, wenn die daran beteiligten Agenten über Vertrauen operieren?

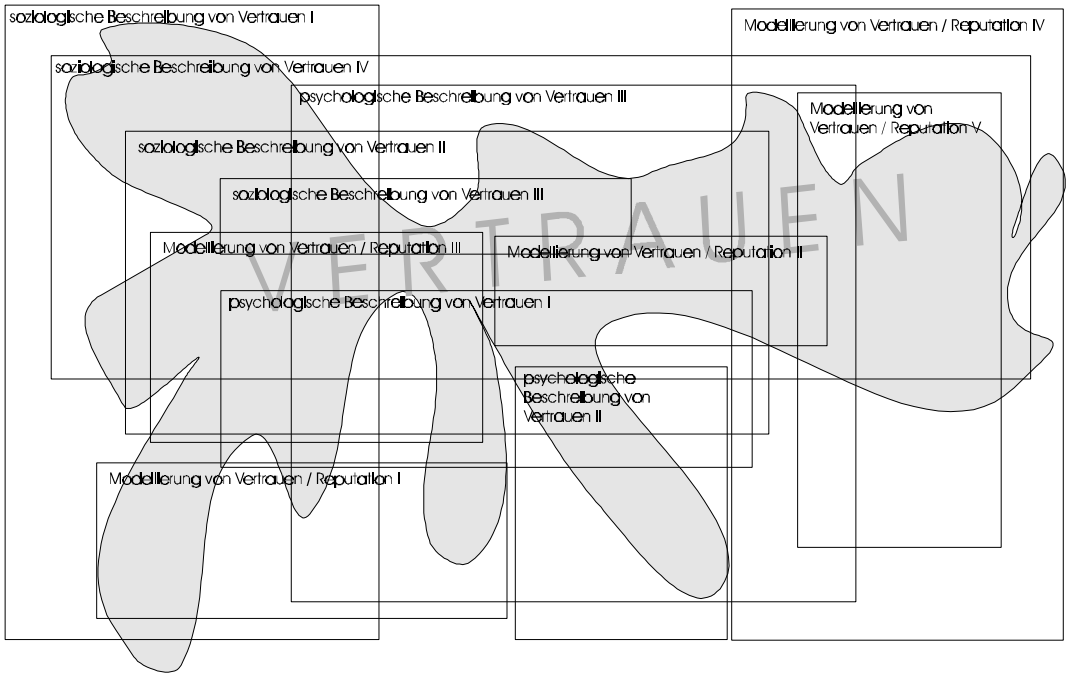


Abb. 1 - Warum es nicht ganz einfach ist, sich komplexen Phänomenen zu nähern

Aufbau der Arbeit

In der Informatik gibt es inzwischen verschiedene Ansätze, die Anregungen aus der Soziologie aufgenommen haben, oder dies zumindest behaupten, und Lösungsmöglichkeiten für das Vertrauensproblem anbieten. Im Rahmen dieser Arbeit soll es nun darum gehen, einige dieser Ansätze zu diskutieren und miteinander zu vergleichen. Dazu wird untersucht, wie die Soziologie Vertrauen theoretisch konzipiert, wie also angenommen wird, dass die hier als Vorbild genommenen Prozesse der Vertrauensbildung real ablaufen (Kapitel 2). Aufbauend darauf wird es darum gehen, die Bedingungen zu klären, unter denen eine Übernahme von soziologisch motivierten Modellen in Multiagenten-Systeme sinnvoll ist, und Kriterien dafür zu entwickeln (Kapitel 3), die dann wiederum als Leitfaden für die vergleichende Diskussion verschiedener Ansätze zur Modellierung von Vertrauen aus dem Bereich der MAS-Forschung, der Internet-Sicherheit sowie der *Distributed Artificial Intelligence* (DAI) dienen kann (Kapitel 4). Das Ziel der Arbeit ist es, Hinweise darauf zu geben, welchen Beitrag diese Ansätze dazu leisten können, einen Schutz vor betrügerischem Verhalten im hier als Anwendungsbeispiel ausgewählten Multiagenten-System AVALANCHE (vgl. dazu Sackmann 1998; Eymann et al. 1998; Eymann / Padovan 1999; Padovan et al. 2000) zu implementieren (Kapitel 5).

2 Das soziologische Vorbild

Vertrauen als Thema

Das Ziel dieses Kapitels ist es, die Funktionen, Rahmenbedingungen und Mechanismen der Vertrauensbildung aus soziologischer Perspektive darzustellen. Obwohl Soziologen und Soziologinnen sich nicht erst seit kurzen, sondern seit der Gründung der Disziplin mit der Frage beschäftigen, welche Rolle Vertrauen etwa für die Aufrechterhaltung und Bereitstellung sozialer Ordnung hat, stand die 'Vertrauensfrage' doch nie im Mittelpunkt der Disziplin: »The concept of trust entered sociological theory by way of philosophical and political writings, never having been a central focus of sociological theory. [… T]rust was seldom explicitly questioned or studied.« (Misztal 1996: 1). Dementsprechend entstanden sehr unterschiedliche Ansätze, wie mit Vertrauen als Element umfassenderer Theorien umzugehen ist, und dementsprechend unterschiedlich sehen auch die Arten des Zugriffs auf das Phänomen Vertrauen aus, die die einzelnen AutorInnen vornehmen (vgl. Abb. 2). Da es hier nicht darum geht, alle soziologischen Zugriffe auf Vertrauen genau darzustellen, sondern vielmehr einige Grundannahmen herausgestellt werden sollen, konzentrieren wir uns im folgenden auf drei unterschiedliche Ansätze. Dazu werden die Analyse von Vertrauen im Rahmen der Systemtheorie durch Niklas Luhmann (1989), der eher umfassend gesellschaftsphilosophisch orientierten Zugriff, der hier durch Barbara Misztal (1996) und Piotr Sztompka (1999) vertreten ist, sowie der eher handlungszentrierten, der Rational-Choice-Theorie nahestehenden Zugang, für den stellvertretend Diego Gambetta (1988b) steht, herangezogen. Die Ansätze von Anthony Giddens (Vertrauen als Teil der Strukturierungstheorie; vgl. dazu Misztal 1996) und James Coleman (Vertrauen im Rahmen der *Rational-Choice-Theory;* vgl. dazu auch Junge 1998; Kollock 1994) können ebenso wie die ökonomisch ausgerichteten Zugriffe auf Vertrauen von Partha Dasgupta und Tanja Ripperger und der sozialpsychologische Zugang von Morton Deutsch hier nicht oder nur am Rande berücksichtigt werden. Verwiesen sei weiter auf die Arbeiten von Bernard Barber, Ulrich Beck, Shmuel Eisenstadt und Francis Fukuyama (vgl. dazu Misztal 1996 und Sztompka 1999). Neben den genannten theoretischen Zugriffen auf Vertrauen existieren natürlich auch auf spezifische Problemstellungen bezogene Ansätze – genannt sei Rainer Kuhlen (1999a, 1999b), der zur Untersuchung der Vertrauenswürdigkeit von 'Informationsassistenten' in Bezug auf elektronische Märkte auf soziologische Vertrauenskonzepte zurückgreift.

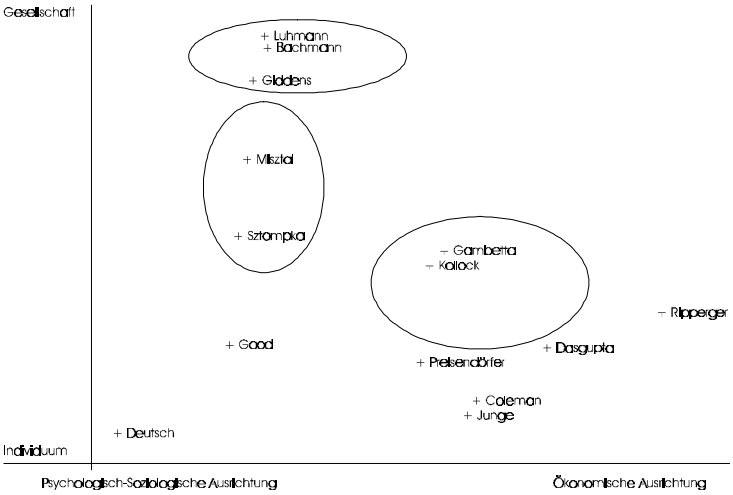


Abb. 2 – Verschiedene Ansätze zur Analyse von Vertrauen[1]

[1] Natürlich kann bei jeder in das Achsenschema Individuum–Gesellschaft und Soziologie/Psychologie–Ökonomie eingeordneten Theorie darüber gestritten werden, ob sie an der richtigen Stelle steht.

2.1 Das Vertrauen der Gesellschaft

Luhmann 1989

Niklas Luhmanns zuerst 1968 erschienene Abhandlung zum Thema *Vertrauen* (hier: Luhmann 1989; vgl. dazu auch Bachmann 1998) wird – zurecht – immer wieder zitiert, wenn es um dieses Thema geht. Der 1998 gestorbene Luhmann, der sich in der Soziologie insbesondere als Systemtheoretiker (vgl. Luhmann 1998) einen Namen gemacht hat, betrachtet Vertrauen vor allem im Hinblick auf die Funktion der Reduktion von sozialer Komplexität. Er unterscheidet dabei zwischen Vertrautheit und Vertrauen, zwischen Vertrauen und Hoffnung, und natürlich zwischen Vertrauen und Misstrauen.

Vertrautheit

»Vertrautheit ist Voraussetzung für Vertrauen wie für Mißtrauen, das heißt für jede Art des Sichengagierens in eine bestimmte Einstellung zur Zukunft.« (Luhmann 1989: 19). Vertrautheit ist damit das latent bleibende Hintergrundwissen über »Lebenswelt, Natur und menschliche Beziehungen« (1989: 22), der Entzug der extremen Komplexität der Welt, und ermöglicht ein »relativ sicheres Erwarten« (1989: 19). Erst vor dem Hintergrund der Vertrautheit ist Vertrauen möglich.

Was ist Vertrauen?

Von der bloßen Hoffnung unterscheidet Luhmann es dadurch, dass die vertrauensvolle Erwartung bei der Entscheidung entscheidend sein muss. »Vertrauen bezieht sich also stets auf eine kritische Alternative, in der der Schaden beim Vertrauensbruch größer sein kann als der Vorteil, der aus dem Vertrauenserweis gezogen wird.« (Luhmann 1989: 24). Dabei geht Luhmann nicht davon aus, dass Vertrauen bereits im Augenblick der Vertrauenserteilung rational berechnet wird. Vielmehr steht das Wissen, ob es sich nachträglich als gerechtfertigt erweist oder nicht, dem Entscheidenden zum Entscheidungszeitpunkt nicht zur Verfügung – »auch nicht in Form bestimmter Wahrscheinlichkeitsziffern« (Luhmann 1989: 25). Gerade durch diese Entscheidung bei fehlendem Wissen – Luhmann spricht vom »Überziehen der vorhandenen Informationen« (1989: 26) – ermöglicht Vertrauen die wechselseitige und in die Zeit gestreckte Verteilung von Komplexität zwischen mehreren Handelnden, und damit die Reduktion von Komplexität. Dadurch wird zumindest eine indifferente Teilnahme an sozialen Handlungen möglich, ohne in jedem Augenblick Chaos bzw. das Schlimmste befürchten zu müssen.

Vertrauen herstellen

Eine wichtige Eigenschaft von Vertrauen, also letztlich dem Ersetzen von (nicht gegebener) äußerer Sicherheit durch innere Sicherheit, einer Verschiebung des Risikos, ist die Tatsache, dass »Menschen und soziale Einrichtungen, denen man vertraut […] besonders störempfindlich sind und gleichsam jedes Ereignis unter dem Gesichtspunkt der Vertrauensfrage registrieren.« (Luhmann 1989: 30). Die Herstellung von Vertrauen ist ein eher langsamer Prozess, der in der wiederholten, riskanten, wechselseitigen Vorausleistung und Aufteilung von Komplexität begründet ist, während jede Störung sehr schnell als solche interpretiert wird und zum Verlust von Vertrauen führt. Die Vereinfachung des Umweltbildes wird mit dieser spezifischen ‘Zerbrechlichkeit’ des Vertrauens erkauft. Schon darin

zeigt sich, wie stark Vertrauen auf Informationen über das Objekt des Vertrauens angewiesen ist.

Indizien für Vertrauen

Sind die Informationen vollständig, ist kein Vertrauen notwendig; liegen keinerlei Informationen vor, ist Vertrauen nur in pathologischer Form möglich (Luhmann 1989: 34). Obwohl Vertrauen immer 'überzogene Information' darstellt, ist ohne Information kein sinnvolles Vertrauen möglich. Vertrauensbildung ist also stark darauf angewiesen, Indizien für Vertrauen zu finden. Luhmann nennt hier insbesondere die Vertrautheit mit dem Objekt des Vertrauens, und die Überlegung, wie die Motivationsstruktur des Partners aussehen mag (1989: 35), aber auch die Funktion latent vorhandener Sanktionsmöglichkeiten wie etwa einer Rechtsordnung, die »das Risiko der Vertrauensgewähr« (1989: 35) entlastet, solange sie nicht direkt angesprochen wird und damit aus einem Vertrauensverhältnis ein juristisches Verhältnis macht. Diese latente Funktion wird durch das »Gesetz des Wiedersehens« (1989: 39) ermöglicht, also der Tatsache, dass Vertrauen wiederholt zwischen den gleichen Personen existiert.

Persönliches Vertrauen

»Man kann Vertrauen nicht verlangen. Es will geschenkt und angenommen sein.« (Luhmann 1989: 46). Persönliches Vertrauen ist nur in Situationen möglich, in denen der Vertrauende auf seinen Partner angewiesen ist. Das Vertrauen muss enttäuscht werden können, und der Vertrauende muss in 'riskante Vorleistung' treten, während der Partner diese Vorleistung honorieren muss, und andere Interessen zurückstellt. Dieses Element der Vertrauensbildung wiederholt sich, mit wachsendem Einsatz. Allerdings darf dieser gegenseitige Lernprozess nicht als kontinuierliche Ausdehnung von Vertrauen gedacht werden, sondern stößt auf Schwellen, deren Überschreitung qualitativ differenziert. (1989: 45ff). Zugleich beruht Vertrauen auf bestimmten 'systeminternen' Voraussetzungen, etwa einer gewissen Selbstsicherheit, die wiederum u.a. auf dem Verhältnis von Kind und Eltern beruht (1989: 85ff, 90). Interessant ist hierbei noch die Anmerkung, dass die für den Prozess der Vertrauensbildung notwendige Selbstdarstellung als vertrauenswürdige Person über einen längeren Zeitraum dazu führt, dass eine Person – so sie nicht den Ort spurlos verlassen möchte – sich auch vertrauenswürdig verhält (1989: 69-71).

Systemvertrauen

Neben diesem auf persönlichen Kontakt und Wiederholungsabsicht angewiesenen Formen sieht Luhmann weitere Formen des Vertrauens. Diese bauen auf standardisierten Normalitäten von Situationen auf und laufen sehr viel anonymer ab. Dazu zählt als Gegenpol zum persönlichen Vertrauen insbesondere das Systemvertrauen, etwa in Geld als generalisiertes Kommunikationsmedium. Mit Geld wird dem Einzelnen ein Ausschnitt aus der Komplexität des Wirtschaftssystem »buchstäblich in die Hand gegeben« (Luhmann 1989: 53); ein Mechanismus, der nur funktionieren kann, wenn dem Geld auf einer sehr abstrakten Basis Vertrauen entgegengebracht wird. Ähnliches gilt für die Medien der Wahrheit und der Macht. Wie auch die Vertrautheit in die Lebenswelt ist das Systemvertrauen etwas, das im Alltag eher latent bleibt. Für die heutige Gesellschaft hat Systemvertrauen eine viel größere

Bedeutung als persönliches Vertrauen – eine Ursache dafür sieht Luhmann übrigens in der Computertechnologie und den durch sie mitverursachten unpersönlichen Kommunikationen (vgl. Luhmann 1998: 312f).

Misstrauen

Zum Verhältnis von Vertrauen zu Misstrauen betont Luhmann, dass Misstrauen »nicht nur das Gegenteil von Vertrauen, sondern als solches zugleich ein funktionales Äquivalent für Vertrauen« (1989: 78) darstellt, da auch Misstrauen zur (oft drastischen) Reduktion von sozialer Komplexität führt. Bedeutsam sind dabei die persönlich je unterschiedlichen Schwellen, bei denen es zu Umschlägen zwischen Vertrauen, Vertrautheit und Misstrauen kommt. Er nimmt an, dass Misstrauen die Tendenz hat, sich als Regelkreis zu verstärken, was nur durch eine Intervention des sozialen Systems begrenzt werden kann (Luhmann 1989: 82, 84). Zur Frage der Rationalität von Vertrauen und Misstrauen meint Luhmann, dass »[o]hne Vertrauen sind nur sehr einfache […] Kooperation möglich [sind]« (1989: 98). Zugleich mit einem steigenden Bedarf an Vertrauen steigt auch der Bedarf von Misstrauen in einem System – bis hin zu institutionalisierten Formen der Kontrolle oder spezifischen Misstrauens-Rollen (1989: 99, 104). Zum Verhältnis von Vertrauen und Misstrauen hält Luhmann (1989: 99) fest, dass Vertrauen viel leichter in Misstrauen verwandelbar ist als andersherum. Sowohl Vertrauen als auch Misstrauen können systemspezifisch oder programmspezifisch festgelegt sein; in diesem Sinne können Systemgrenzen Schwellen der Vertrauensänderungen sein (1989: 102).

Fazit

Vertrauen (und Misstrauen) hat bei Luhmann also sowohl im persönlichen Handlungsbereich als auch im Bereich von größeren Systemen die Funktion, Komplexität zu reduzieren, die Welt zu vereinfachen, und so Handlungen effizienter möglich zu machen. Vertrauen ist nicht der einzige soziale Mechanismus, der diese Funktion erfüllt, und kann dies auch nicht alleine, hat aber doch eine fundamentale Bedeutung dafür.

2.2 Gesellschaftlicher Zusammenhalt durch Vertrauen

Misztal 1996

Barbara Misztal (1996) beschreibt in ihrem Buch *Trust in Modern Societies* die gesellschaftlichen Funktionen und verschiedenen Formen von Vertrauen beschreibt. Sie geht dabei ausführlich auf die meisten anderen Autoren ein, die direkt oder – im Bereich der Klassiker – im Zusammenhang mit der Diskussion über soziale Ordnung Beiträge zur Erforschung von Vertrauen geliefert haben. Ihr eigenes Interesse liegt dabei vor allem auf der Rolle, die Vertrauen als Voraussetzung für den gesellschaftlichen (politischen) Zusammenhalt spielt, wobei sie nicht nur die soziologischen Theorien und Definitionen der Klassiker und der modernen SoziologInnen heranzieht, sondern sich auch auf empirische Fallbeispiele beruft. Vertrauen ist für Misztal letztlich eine notwendige Bedingung für das Wohlergehen, die Stabilität und das Vorhandensein von Kooperation:

> The main aim of this book has been to support the belief in the necessity of rethinking social relationships in a more cooperative mood. Evidence of the importance of trust to

> the well-being and stability of society suggests that to achieve a new quality of compliance – that is, social cooperation – we need to devote more attention to the relationships among people and between people and decisions makers. It can be concluded that if it is our responsible conduct and trust that holds us together, the ongoing process of global interdependency will only increase the demand for trust as an essential condition of cooperation. (Misztal 1996: 269)

Formen von Vertrauen

Im Rahmen dieser Arbeit ist Misztal deswegen besonders interessant, weil sie zu einer eigenständigen Klassifizierung von Vertrauen kommt, die deutlich macht, wie vielfältig und weitreichend diese Problemstellung ist. Sie unterscheidet drei verschiedene Formen sozialer Ordnung, für die Vertrauen jeweils eine spezifische Funktion erfüllt: *Stabilität* umfasst die Fragen der Zuverlässigkeit und Vorhersagbarkeit des Sozialen, *Kohäsion*, die auf normativer Integration basiert, und schließlich *Kollaboration*, die sich auf Fragen der sozialen Kooperation bezieht. (Misztal 1996: 64). Der Stabilitätsaspekt sozialer Ordnung wird von Misztal mit einem habituellen Konzept von Vertrauen (*Habitus*) verbunden, das beschrieben wird als »routine background to everyday interaction through which the predictability, legibility and reliability of collective order is sustained, while the perception of its complexity and uncertainty is restricted.« (1996: 97). Unter dem kohäsiven Aspekt sozialer Ordnung verknüpft sie Vertrauen mit Familiarität, freundschaftlichen Bindungen sowie gemeinsamen Werthaltungen (*Passion*), während in Bezug auf Kollaboration Vertrauen als eine gesellschaftlich-politische Haltung beschrieben wird, die daraufhin zielt, Solidarität, Toleranz und gegenseitigen Respekt zu ermöglichen und unkooperative Einstellungen abzuschwächen (*Policy*). Jede dieser Formen von Vertrauen erfüllt nicht nur bestimmte gesellschaftliche Funktionen, sondern hängt auch mit bestimmten Formen der gesellschaftlichen Praxis zusammen (vgl. Tabelle 1).

Order	Trust	Practice
Stable	Habitus	Habit, Reputation, Memory
Cohesive	Passion	Family, Friends, Society
Collaborative	Policy	Solidarity, Toleration, Legitimacy

Tabelle 1 – Formen des Vertrauens nach Misztal (1996: 101)

Reputation und mehr

Auch wenn Misztals Ansatz, was nicht nur in ihrem Fazit deutlich wird, an einigen Stellen doch stark normativ gefärbt ist, macht ihr breites Vorgehen deutlich, dass Vertrauen deutlich mehr ist als nur ein Instrument, um die Zusammenarbeit zwischen zwei Parteien herzustellen, und zwischen erwünschten und unerwünschten Kooperationspartnern zu trennen. Vertrauen ist auch das, wie Misztal (1996: 120ff) unter dem Gesichtspunkt der *Reputation* erörtert. Sie geht dabei unter anderem auf die Bedingungen für ökonomische Beziehungen ein (vgl. Gambetta 1998b), geht aber weit darüber hinaus. Neben den Mechanismus der interessengeleiteten Reputationsbildung aus einem Nutzenkalkül heraus stellt sie die Reputation, die dadurch erzeugt wird, das eine Person einer Gruppe beitritt,

der Ehrerbietung (in einem weiteren Sinn) entgegengebracht wird, also eine ethisch-moralisch begründete, auf Werte setzende Form der Reputation. In diesem Sinn kann Reputation auch durch Disziplin gefördert werden. Obwohl heute nicht mehr der Ehrenkodex mittelalterlicher Gesellschaften herrscht, ist doch die Mitgliedschaft in bestimmten Professionen oder Organisationen noch immer reputationsvergrößernd. Allerdings sieht Misztal hier auch starke Gegentendenzen – ironischerweise etwa durch die Anonymität elektronischen, nicht mehr sozial eingebetteten Einkaufens (1996: 134). Eine andere Art der gruppenspezifischen Reputation (hier eher negativ betrachtet) sind zum einen die Vorurteile, die etwa MigrantInnen-Populationen entgegengebracht werden, positiv gewendet aber auch die aus der Not heraus geborene vorausgesetzte gegenseitige Anerkennung innerhalb solcher Populationen. Allgemein beschreibt sie Reputation »as part of a simplifying process in which the unknown and fearful becomes familiar, singles out individuals or groups noted for distinction, respectability or good name.« (Misztal 1996: 123). Wichtig ist vielleicht noch, darauf hinzuweisen, dass Reputation von Vertrauen in einem engeren Sinne – hier wohl eher im Sinne von Vertrautheit gebraucht – abgegrenzt wird (Misztal 1996: 121). Um es noch einmal zu betonen: Dieses vielfältige Bild davon, was Reputation alles sein kann, und wie Reputation beeinflusst wird, ist für Misztal nur eines von neun Elementen, die als unterschiedlichen Praxen alle zusammen unter dem Oberbegriff Vertrauen behandelt werden.

Sztompka 1999

Ähnlich, allerdings nicht ganz so weitreichend, argumentiert Piotr Sztompka (1999) in *Trust. A Sociological Theory* aus einer an Kultur orientierten Perspektive heraus. Seine Motivation liegt dabei vor allem darin, herauszufinden, wie Kulturen des Vertrauens entstehen und zugrunde gehen, und mit diesem analytischen Modell die Besonderheiten der postsozialistischen sozialen Ordnung in Osteuropa zu untersuchen.. Für uns interessant ist vor allem seine Beschreibung verschiedener Formen und Funktionen von Vertrauen, der er die These vorausschickt, dass Vertrauen ein besonderes Merkmal moderner Gesellschaften ist. Ältere Gesellschaften waren auf Vertrauen als Schmiermittel zwischenmenschlicher Handlungen nicht angewiesen (vgl. Sztompka 1999: 11ff). Dies klingt zuerst einmal paradox, wird aber mit folgender Überlegung deutlicher: Wo alles klar geregelt ist – durch Normen, religiöse Vorschriften, starre Hierarchien – ist Vertrauen nicht notwendig. Ebenso sieht es aus, wenn über die Handlungen des Gegenübers volle Kontrolle besteht. Erst, wenn Risiko, Unsicherheit, Unwahrscheinlichkeit ins Spiel kommt – und dies ist beispielsweise dann der Fall, wenn Gesellschaften sich funktional ausdifferenzieren, wenn viele Dinge durch Marktbeziehungen geregelt werden, statt hierarchisch gegliedert zu sein – wird Vertrauen zu einem notwendigen Element der Gesellschaft. Vertrauen hat dann die Funktion hat, Unsicherheit zu reduzieren und mit Risiko umzugehen (vgl. auch Luhmann 1989).

Definition Vertrauen

Sztompka definiert Vertrauen grundlegend wie folgt: »Trust is a bet about the future contingent actions of others« (1999: 25) – er konzipiert also Vertrauen als eine Wette über

die zukünftigen, kontingenten Handlungen anderer. Damit weist er zugleich darauf hin, dass Vertrauen immer eine Eigenschaft gegenüber Menschen bzw. menschlichen Handlungen ist. Einem Vulkan kann nicht vertraut werden, und einer gesellschaftlichen Institution zu vertrauen, heißt immer auch, den Menschen dahinter zu vertrauen (1999: 21, 41ff). Er koppelt seinen Vertrauensbegriff an drei unterschiedlich motivierte Formen des *commitment* (Sztompka 1999: 27ff) – Vertrauen, das dadurch motiviert ist, dass die Handlungen einer anderen Person den eigenen Interessen und Bedürfnissen zuträglich sein wird (*anticipatory trust*); Vertrauen, das durch die vermutete Reaktion der Gegenseite motiviert ist – etwa, wenn es darum geht, ein wertvolles Objekt[2] zeitweise jemand anderem zu überlassen (*responsive trust*); sowie Vertrauen, das dadurch motiviert ist, dass erwartet wird, dass einem selbst Vertrauen entgegengebracht wird (*evocative trust*). Diese unterschiedlichen Formen, Vertrauen auszusprechen, tauchen dazu noch in unterschiedlichen Stärken auf, wobei die Stärke (*degree of commitment*) wiederum abhängig ist von der erwarteten Zeitspanne der Vertrauensbeziehung, von der Möglichkeit, sein Vertrauen zurückzuziehen, von der Größe des mit dem Vertrauen verbundenen Risikos (also den Folgen gebrochenen Vertrauens), vom Vorhandensein von Versicherungen gegen die Verluste aufgrund gebrochenen Vertrauens sowie, wenn es darum geht, anderen ein Objekt zu überlassen, vom subjektiven Wert dieses Objekts. All diese Faktoren verändern Vertrauensbeziehungen; besonders wichtig ist dabei allerdings die Verbindung zwischen Vertrauen und Risiko.

Wem vertrauen?

Wenn es darum geht, ob einer anderen Person Vertrauen – in einer der oben genannten Formen – entgegengebracht werden soll oder nicht, unterscheidet Sztompka verschiedene Hinweisen. Da mit jeder Vertrauenshandlung auch das Risiko eingegangen wird, die oben genannte Wette gegen die zukünftigen Handlungen anderer zu verlieren, ersetzt Vertrauen zwar – gesellschaftlich gesehen – Risiko, setzt an dessen Stelle jedoch wiederum ein persönliches Risiko, nämlich dass, diese Wette zu verlieren. Es müssen also – ähnlich argumentiert, wie wir gesehen haben, auch Luhmann – Hinweise darauf gesucht werden, warum diese Wette eingegangen werden soll. Diese Hinweise auf die Vertrauenswürdigkeit eines anderen können aus der Beziehung zwischen zwei Handelnden kommen, sie können psychologisch motiviert (vgl. dazu auch Good 1998) sein, oder sie können aus dem situativen Kontext stammen. Auch Sztompka geht hier also deutlich über die bloße Kosten-Nutzen-Berechnung des *Rational-Choice*-Ansatz hinaus, in dem situative und psychologische Aspekte meist keine Berücksichtigung finden. Sztompka nennt einige solcher Hinweissysteme: Reputation (hier definiert als die Aufzeichnung vergangener Taten), Performanz (Bewertung der aktuellen Handlungen, nicht der Handlungsgeschichte), Erscheinung (z.B. Kleidung), zukünftige Verfügbarkeit des Verantwortlichen (ein Straßenverkäufer ist

[2] Diese Form des Vertrauens ist immer mit einem spezifischen Objekt verbunden, das anderen überlassen wird, es handelt sich hier also nicht um allgemeines Vertrauen.

zukünftig nicht verfügbar, um zur Verantwortung gezogen werden, die Institution Sotheby's ist es durchaus), *pre-commitment* (d.h., den Kontext absichtlich so ändern, dass die Folgen der Nichtausführung der Handlung risikoreicher werden, vgl. auch Gambetta 1988b), sowie einen psychologisch motivierten *trusting impulse*. (vgl. Sztompka 1999: 69ff).

Rolle der Reputation

Im Kontext dieser Arbeit scheint die Frage der *Reputation* eine besondere Rolle zu spielen. Sztompka definiert Reputation, wie oben beschrieben, als » record of past deeds« (1999: 71). Etwas ausführlicher dargestellt, bezieht sich dies darauf, dass eine Person oder die Institution, der Vertrauen entgegengebracht werden soll, in den meisten Fällen schon seit einiger Zeit existiert. Zum einen ist es so möglich oder sogar wahrscheinlich, dass es früher schon zu Kontakten gekommen ist, die jetzt ausgewertet werden können. Zum anderen ist zu vermuten, dass es Informationen über diese Person oder Institution gibt, die von Dritten stammen, die eigene Erfahrungen mit dem Objekt des Vertrauens gemacht haben. Auch Dinge wie Zeitungsartikel, Zeugnisse, Medaillen, usw. Hinweise auf die Vergangenheit des zu vertrauenden Menschen oder der zu vertrauenden Institution.[3] Die Aufzeichnung vergangener Taten muss allerdings nicht einmal materiell vorhanden sein – »word of mouth is the most common mode in which such records are kept«, schreibt Partha Dasgupta (1988: 66, Fn. 19) in einem ähnlichen Zusammenhang.

Fazit

Zusammenfassend lässt sich sagen, dass Menschen, wenn sie darüber entscheiden, ob sie anderen Menschen oder Institutionen vertrauen, sich von einer Vielzahl an relationalen, kontextbezogenen und auch psychologischen Hinweisen leiten lassen. Reputation ist ein besonders wichtiger Bestandteil dieses Hinweisbündels, meint aber zugleich sehr breit gefasst alle Informationen, die direkt oder indirekt über das vergangene Handeln einer Person oder Institution in Erfahrung zu bringen sind. Dazu gehört mit Misztal auch die Zugehörigkeit einer Person zu einer Organisation mit Reputation.

2.3 Vertrauen und kooperatives Handeln

Gambetta 1988

Diego Gambetta (1988b) beschäftigt sich in der Zusammenfassung des von ihm herausgegebenen Sammelbands *Trust. Making and Breaking Cooperative Relations* (Gambetta 1988a) unter dem Titel *Can We Trust Trust?* vor allem mit dem Zusammenhang von Vertrauen einerseits und kooperativem Handeln andererseits aussieht. Ist Kooperation ohne Vertrauen möglich? Gambettas Perspektive ist dabei insbesondere auf das Handeln und die Entscheidungslage einzelner Akteure gerichtet, unter anderem greift er dabei auch auf Ergebnisse der Spieltheorie zurück.

---

[3] Auch Sztompka geht in diesem Zusammenhang auf die besondere Rolle von Hinweisen auf die Verlässlichkeit anderer im Bezug auf das viele klassische derartige Hinweise ignorierende Internet ein (1999: 73).

Definition Kooperation

Da es bei Gambetta um den Zusammenhang von Vertrauen und Kooperation geht, liegt es nahe, diese beiden Begriffe erst einmal zu definieren. Unter *Kooperation* versteht Gambetta die Tatsache, dass zwei Akteure[4] explizit oder implizit Regeln vereinbaren, die im Verlauf ihrer Interaktion dann auch eingehalten werden (Gambetta: 1988a: 213, Fn. 2). Kooperation steht Konkurrenz gegenüber; allerdings zeigt Gambetta deutlich, dass weder eine rein auf Kooperation noch eine rein auf Konkurrenz ausgerichtete Gesellschaft viabel wäre, sondern dass Konkurrenz nur existieren kann, wenn ein gewisser Grad von Vertrauen darauf, dass andere Akteure allgemeine Regeln einhalten, besteht. Er macht allerdings auch deutlich, dass Kooperation per se nicht unbedingt etwas positives sein muss; auch innerhalb von Verschwörerkreisen etwa wird – notwendigerweise – kooperiert. Zur Frage, wie wahrscheinlich kooperatives Verhalten ist, verweist Gambetta auf die Ergebnisse der Spieltheorie, deren Ergebnisse insgesamt eine große Unwahrscheinlichkeit kooperativer Gleichgewichtszustände zeigen. (Gambetta 1998b: 214ff).

Definition Vertrauen

Vertrauen definiert er als »a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action […] *and* in a context in which it affects *his own action*.« (Gambetta 1988a: 217, kursiv i. O.). Ein Akteur ist dann *vertrauenswürdig*, wenn die Wahrscheinlichkeit dafür, dass eine Handlung ausgeführt, die dem Akteur nützt oder zumindest nicht schadet, hoch genug ist, um eine Kooperation in Betracht zu ziehen. Dabei bezieht sich Vertrauen nicht auf alle möglichen zukünftigen Handlungen, sondern nur auf die für die anstehende Entscheidung relevanten zukünftigen Handlungen. Eine Wahrscheinlichkeit p von 0,5 für Kooperation bedeutet bei Gambetta, dass unklar ist, ob einem anderen Akteur vertraut werden kann oder nicht, sinkt p unter 0,5, dann liegt Misstrauen vor, steigt p über 0,5, besteht Vertrauen. Dieses bezieht sich bei Gambetta – hier lehnt er sich an Luhmann an – allerdings nicht nur auf die angenommene oder erwartete Wahrscheinlichkeit dafür, dass eine angestrebte Handlung von anderen Akteuren ausgeführt wird, sondern hängt auch stark von einem bestimmten Grad an Freiheit ab. Nur wenn mindestens einer der beteiligten Akteur die Möglichkeit hat, das in ihn gesetzte Vertrauen durch Fehlverhalten zu enttäuschen oder eine riskant erscheinende Beziehung abzulehnen, und wenn zugleich die Handlungsoptionen so eingeschränkt sind, dass diese riskante Beziehung interessant ist, kann davon gesprochen werden, das ein Akteur einem anderen vertraut (Gambetta 1988b: 219).

Zwang und Interesse

Zum Zusammenhang von Kooperation und Vertrauen zählt Gambetta verschiedene Möglichkeiten auf, Kooperationen durch soziale Arrangements zu beeinflussen, die nicht

---

[4] Gambetta spricht im Original von *agents* und meint damit nicht nur Individuen, sondern auch Firmen oder Regierungen. Das *agent* der soziologischen Literatur wird im Deutschen üblicherweise als Akteur (und nicht als Agent) bezeichnet (vgl. Schulz-Schaeffer 1998). Agent soll im Rahmen dieser Arbeit auf technische Agenten beschränkt bleiben.

direkt an Vertrauen gebunden sind. Dazu gehört zum einen das mögliche Ausmaß an Zwang, der die Handlungsoptionen eines Akteurs so einschränken kann, dass die *exit*-Option nicht mehr gegeben ist, und eine Kooperation ohne Vertrauensgrundlage zustande kommt. Gambetta geht allerdings davon aus, dass eine primär auf Vertrauen basierende Gesellschaft sehr viel effizienter ist als eine durch Zwang, Verpflichtungen und Gewalt gekennzeichnete. Die andere Möglichkeit, Kooperationen wahrscheinlicher zu machen, ohne dass Vertrauen vorausgesetzt wird, sieht Gambetta in verschiedenen Formen des *pre-commitment*, also der willentlichen Einschränkung der Handlungsmöglichkeiten[5]. Verträge und Versprechen sind schwächere Formen des *pre-commitments* – beide führen dazu, dass ein Nichteinhalten einer Kooperation teurer wird, was für einen rationalen Akteur gleichbedeutend mit einer Einschränkung der Optionen ist. Damit kommt das *Interesse* eines Akteurs in Spiel – ein Faktor, der letztlich, wenn der erwartete Nutzen einer bestimmten Option nur hoch genug ist, auch dazu führen kann, dass die situationsabhängige Schwelle der Vertrauenswürdigkeit, ab der ein Akteur sich für Kooperation entscheidet, in Extremfällen sogar unterhalb 0,5 sinken kann. Zur Erklärung dieser Fälle bezieht sich Gambetta auf die Theorie kognitiver Dissonanz, die erklärt, warum Menschen in bestimmten Fällen eher ihre Wahrnehmung oder ihre Einstellung ändern, als eine nicht zu ihren übrigen Kognitionen passende Tatsache hinzunehmen. (Gambetta 1988b: 220ff).

Ist Vertrauen sinnvoll?

Schließlich stellt sich Gambetta die für diese Arbeit besonders interessante Frage, ob Vertrauen nicht nur ein Nebeneffekt von Kooperation ist, der im Prinzip ignoriert werden kann. Er begründet dies mit der Tatsache, dass sowohl evolutionstheoretisch als auch aus Sicht der Spieltheorie (vgl. Axelrodt 1991) das Zustandekommen von Kooperation auch durch kleine, zufällige, vielleicht sogar falsch gedeutete anfängliche Signale der Zusammenarbeit erklärt werden kann, die dann zu einer sich selbst verstärkenden Kooperation führen können. Dabei entsteht letztlich möglicherweise auch der gegenseitige Ruf der Vertrauenswürdigkeit (Reputation), die aber eben nicht Ausgangspunkt der Kooperation war. (Gambetta 1988b: 224ff). Dennoch plädiert Gambetta dafür, sich für Vertrauen als gesellschaftliche Grundlage einzusetzen. Er stellt sich die Frage, wie dies am besten möglich ist, da Vertrauen ja nicht durch eine bloße Willenserklärung dazu zustande kommt. Vertrauen kann als *spin-off*-Effekt von Freundschaft, Moral oder Religion angesehen werden, hängt aber genauso sehr von den sozialen, wirtschaftlichen und politischen Umständen ab. Ausgehend davon, dass der Nachweis eines Missbrauchs von Vertrauen leicht fällt, es aber schwierig ist, das Gegenteil nachzuweisen, kommt er zu der Schlussfolgerung, das Vertrauen vom Fehlen konträrer Indizien abhängt, und dadurch auch sehr leicht – mutwillig oder unabsichtlich – zerstört werden kann. Ein Aufbau von Vertrauen

---

[5] Ein Beispiel dafür wäre ein gemeinsames Bankkonto, dass nur bei Zeichnung durch zwei Personen genutzt werden kann – damit besteht nicht mehr die Möglichkeit, dass ein Akteur den anderen ausnützt; die Notwendigkeit für Vertrauen nimmt also ab.

aus der Situation des Misstrauens heraus ist dabei fast unmöglich. Das Rezept, um Vertrauen zu bilden, heißt deswegen für ihn: »trust begins with keeping oneself open to evidence, acting *as if* one trusted, at least until more stable beliefs can be established on the basis of further information.« (Gambetta 1988b: 234). Daraus lässt sich ableiten, dass die Zunahme von Vertrauenswürdigkeit über die Zeit in relativ kleinen Schritten erfolgt – und zwar unabhängig von speziellen Vertrauensbeweisen –, während die Abnahme aufgrund konträrer Evidenz recht schnell vonstatten geht. In beiden Fällen stellen die Werte 1,0 bzw. 0,0 Grenzen dar, die nicht überschritten werden können – mehr Vertrauen als immer zu vertrauen und weniger Vertrauen als immer zu misstrauen gibt es nicht.

2.4 Zusammenfassung

Fazit

In diesem Kapitel wurden die soziologischen Ansätze von Luhmann, Misztal, Sztompka und Gambetta dargestellt. Luhmann stellt Vertrauen in den Kontext der Systemtheorie. Vertrauen wird auf Risiko bezogen, seine Hauptfunktion liegt in der Reduktion von Komplexität. Er stellt u.a. dar, wie Prozesse der Vertrauensbildung ablaufen und welche Bedingungen – insbesondere ein Mittelding zwischen Wissen und Nichtwissen – für Vertrauen gegeben sein müssen. Luhmann betont, dass Vertrauensbildung langsam und die Zerstörung von Vertrauen schnell abläuft, und dass es qualitative Schwellen gibt. Misztal benutzt einen sehr viel umfassenderen Begriff von Vertrauen als die anderen Autoren. Sie bezieht Vertrauen auf den Erhalt der sozialen Ordnung. Bei ihr umfasst Reputation als eine von neun Praxisformen des Vertrauens nicht nur die auf vergangenes Verhalten bezogene Reputation, sondern auch Reputation aufgrund von Gruppenzugehörigkeit. Sztompka definiert Vertrauen als Wette mit der Zukunft. Verschiedene von ihm unterschiedene Formen des Vertrauens und Hinweissysteme, wem zu vertrauen ist, werden geschildert, insbesondere Reputation, hier im Sinne einer – auch mündlichen – Aufzeichnung vergangenen Verhaltens. Gambetta besitzt einen der Ökonomie bzw. der *Rational-Choice*-Theorie nahestehenden Vertrauensbegriff. Er untersucht insbesondere das Verhältnis von Kooperation zu Vertrauen und betont insbesondere, dass Kooperation auch ohne Vertrauen möglich ist. Daraus ableitend gibt er Hinweise darauf, wie Vertrauen in einem wechselseitigen Prozess gebildet wird. Auch Gambetta unterscheidet zwischen unterschiedlichen Geschwindigkeiten der Vertrauensbildung und der Misstrauensbildung. Ohne zu einem einheitlichen Begriff zu kommen, machen die verschiedenen TheoretikerInnen deutlich, welche Funktionen Vertrauen hat, wie es gebildet wird und wo seine Grenzen liegen.

3 Zum Zusammenhang von Theorie und Praxis

3.1 Ist Vertrauen auf Agentensysteme anwendbar?

Es stellt sich nun die Frage, wieweit Vertrauen auf technische Systeme, speziell auf Multiagenten-Systeme, übertragbar ist. Bisher spielte Vertrauen in Bezug auf technische Systeme zwar immer dann eine große Rolle, wenn es um das Vertrauen von Menschen *in*

technische Systeme ging – also etwa bei Fragestellungen, ob einem Flugzeug, einem Kernkraftwerk oder einem Auto vertraut werden kann (vgl. dazu Kuhlen 1999b; vgl. auch Luhmann 1998), nicht aber im Bezug auf die Zusammenhänge innerhalb technischer Systeme: Es ist für eine Grafikkarte nicht notwendig, der CPU, dem Bildschirm oder der Treibersoftware zu vertrauen. Natürlich war es auch bisher schon möglich, die Zusammenhänge innerhalb eines komplexeren technischen Systems – metaphorisch, anthropomorph – mit Begriffen wie Vertrauen zu beschreiben. Sinnvoll wird diese Begrifflichkeit allerdings erst dann, wenn bestimmte Voraussetzungen gegeben sind, die hier abgeklärt werden sollen.

Entscheidende Akteure

Die wichtigste dieser Voraussetzungen ist vielleicht die Tatsache, dass das Konzept Vertrauen nur dann sinnvoll anwendbar ist, wenn zumindest eine Seite – eher noch beide[6] – in der Vertrauensbeziehung dazu in der Lage ist, Entscheidungen zu treffen und diesen Entscheidungen zufolge zu handeln. Peter Preisendörfer (1995: 264) spricht deswegen auch von Vertrauen als einem Merkmal sozialer Beziehungen, in die jeweils mindestens zwei Akteure involviert sind.

Mittlere Informiertheit

Eine zweite Voraussetzung liegt darin, dass Vertrauen nur da einen Sinn ergibt ist, wo es um zukünftige Handlungen geht, die nicht vollständig kontrollierbar sind. Wenn wir vertrauen, gehen wir eine Wette über die zukünftigen, kontingenten Handlungen anderer ein (Sztompka 1999: 25), d.h., wir handeln unter Unsicherheit bzw. unter Risiko, und wissen eben nicht genau, ob das Vertrauen, das wir in andere setzen, auch erfüllt wird, oder ob unser Handeln uns letztlich schadet (vgl. hierzu insbesondere auch Preisendörfer 1995). Anders gesagt: Vertrauen muss auch enttäuscht werden können (z.B. Luhmann 1989: 45; Gambetta 1988b: 218f). Wenn wir allerdings genau wüssten, ob es enttäuscht wird oder nicht, müssten wir nicht vertrauen, sondern könnten eine sichere Entscheidung treffen. Und wenn wir gar nichts wissen, ist Vertrauen eine ‘blinde’ Vorleistung. Daraus lässt sich ableiten, dass Vertrauen in einer Umgebung, in der alles tatsächlich kontrolliert wird, nicht sinnvoll und im Prinzip auch nicht möglich ist, genau so wenig, wie in einer Umgebung, in der keinerlei Information verfügbar ist.

Gesetz des Wiedersehens

Zum anderen lässt aus dieser Tatsache aber auch schlussfolgern – eine dritte Voraussetzung –, dass Vertrauen etwas ist, das letztlich nur dann etabliert werden kann, wenn zumindest die Möglichkeit wiederkehrender Beziehungen besteht. Ein Punkt, auf den u.a. Luhmann (1989: 39) hinweist, und der auch in David Goods sozialpsychologischer

---

[6] Nach Sztompka (1999: 18ff, 41ff) *müssen* sogar beide Seite dieser Beziehung handelnde Personen sein, bzw. zumindest soziale oder sozio-technische Gebilde, die letztlich handelnden Personen zugerechnet werden können, etwa wenn es um das Vertrauen in eine Fluglinie, ein technisches Artefakt oder gar um das (systemische) Vertrauen in eine Gesellschaft geht. Auch Gambetta beschränkt die Diskussion um Vertrauen auf »trust between agents and excludes that between agents and natural events.« (1988b: 218), wobei mit *agents* hier ebenfalls Personen oder soziale Gebilde wie Firmen oder Regierungen sind..

Untersuchung zu den Bedingungen von Vertrauen und Kooperation eine wichtige Rolle einnimmt (Good 1988: 37).

Vertrauen bei Agenten?

Ausgehend von den drei genannten Voraussetzungen – handelnde, entscheidungsfähige Akteure; ein Mittleres zwischen Unwissenheit und vollständiger Kontrolle; die Möglichkeit wiederkehrender Beziehungen – geht es jetzt darum, die Übertragbarkeit eines Vertrauenskonzepts auf Agentensysteme näher zu bestimmen. Der größte Teil der soziologischen Literatur geht davon aus, dass nur Menschen – und abgeleitet davon vielleicht soziale Organisationen und Gebilde – handelnde, entscheidungsfähige Akteure sind, dass dies aber nicht auf technische Artefakte zutrifft. Diese grundlegende Unterscheidung wird von der Informatik mit dem Begriff autonomer Agenten in Frage gestellt. Im Rahmen dieser Arbeit gehen wir davon aus, dass diese Erweiterung des Handlungsbegriffs auf technische Artefakte tragbar und sinnvoll ist, obwohl dies natürlich umstritten ist (vgl. dazu etwa Schulz-Schaeffer 1998; Braun 1998). Wir vermuten also, dass hinreichend autonome Agenten im technischen Sinne die erste Voraussetzung für die Anwendbarkeit von Vertrauen erfüllen. Dazu müssen diese Agenten allerdings in der Lage sein, Entscheidungen zu treffen, die andere Agenten oder Systeme nicht von vorneherein wissen können.

Wie sieht es mit der zweiten Vorraussetzung aus – dem mittleren Informationsstand zwischen Unwissenheit und totaler Kontrolle? Hier scheint mir sehr viel davon abzuhängen, welcher Anwendungsbereich gewählt wird. In vielen geschlossenen Informationssystemen ist das Verhalten einzelner Agenten prinzipiell – etwa durch eine zentrale Steuerungssoftware – nicht nur vorhersagbar, sondern einsehbar. In einem solchen, geschlossenen System, in dem Vertrauen nicht enttäuscht werden kann, scheint es daher nur dann sinnvoll zu sein, auf Vertrauen zu setzen, wenn sich dadurch Vorteile in der Effizienz ergeben, oder wenn es sich um ein System handelt, dass speziell dazu dienen soll, Entscheidungen unter Unsicherheit oder unter Risiko zu simulieren; wenn also bewusst vorhandene Informationen nicht weitergegeben werden, sondern erraten und geschätzt werden müssen und – Vertrauen unter totaler Unwissenheit ist pathologisch – auch zu einem gewissen Grad erraten oder geschätzt werden können. Anders sieht es aus, wenn es sich um ein offenes Informationssystem handelt, wenn also beispielsweise Agenten ausgeführt werden können, deren Intention (und ‘Vertrauenswürdigkeit’) per se weder dem System noch den anderen Agenten bekannt ist, oder wenn es um Interaktionen zwischen Menschen und technischen Agenten geht.

Die dritte Voraussetzung schließlich besteht darin, dass zumindest potenziell wiederkehrende Beziehungen möglich sind. Das bedeutet nicht nur, dass ein Agent mehr als einmal mit einem anderen Agenten interagiert, sondern vielmehr auch, dass diese Agenten dabei die Möglichkeit haben, auf *Wissen* über ihre früheren Interaktion zurückzugreifen. Hinter dieser so einfach aussehenden Voraussetzung lauern recht schwerwiegende Anforderungen an die Identifizierbarkeit der Agenten über längere Zeiträume, und gewissermaßen auch an

die ‘Kohärenz’ eines Agenten, d.h. daran, dass das Verhalten eines Agenten sich im Allgemeinen nicht zu abrupt verändert, sondern stabil bleibt. Außerdem führt diese Voraussetzung dazu, dass Agenten zumindest rudimentär über ein Gedächtnis für Interaktionen mit anderen Agenten verfügen müssen, und bezüglich ihres Verhaltens auf dieses Gedächtnis zurückgreifen können müssen.

Fazit

Wenn wir diesen drei Voraussetzungen folgen möchten, erscheint es demzufolge nicht sinnvoll, von Vertrauen etwa zwischen der oben erwähnten Grafikkarte (gedacht als einem autonomen Agenten) und ihrer Treibersoftware zu sprechen. Die Interaktionsmöglichkeiten und Entscheidungsfreiheiten sind hier zu eingeschränkt – das Versagen der Grafikkarte oder des Treibers, das gewünscht zu tun, wäre ein technischer Defekt, und nicht die mutwillige Enttäuschung des wechselseitigen Vertrauens. Deutlich wird auch, dass sich nicht jedes Multiagenten-System automatisch für die Implementierung von Vertrauen bzw. von der sozialen Funktion von Vertrauen äquivalenten Mechanismen eignet. Um es zuzuspitzen: Nur solche Multiagenten-Systeme, in denen die insbesondere in der *Distributed Artificial Intelligence*-Forschung häufig gemachte Vorstellung des gutwilligen Agenten nicht gültig ist, in denen also Agenten sich betrügerisch[7] verhalten können – sei es zu Simulationszwecken, sei es, weil betrügerisches Verhalten auf einem offenen elektronischen Marktplatz Gewinn verspricht – eignen sich überhaupt möglicherweise dafür, Vertrauen zu implementieren.

3.2 Zwischen Simulation und technischer Implementierung

Kognitive Perspektive

Allerdings stellt sich auch in den so eben geschilderten Fällen die Frage, wieweit prinzipiell soziale Beziehungen in technische Systeme übertragbar sind. Mit Good (1988; vgl. Mueller 1995; vgl. Marsh 1994a: 142) bin ich der Meinung, dass die hinter vielen Formalisierungen von Vertrauenswürdigkeit stehende Annahme unbeschränkter Rationalität der Akteure nicht haltbar ist. Ein wirkliches Verständnis der Entwicklung und Aufrechterhaltung von Vertrauensbeziehungen scheint nur dann möglich zu sein, wenn berücksichtigt wird, dass aus einer kognitiven Perspektive der menschlichen Informationsverarbeitung die Gehirne von Menschen nicht in der Lage dazu sind, vor dem Eingehen von Vertrauensbeziehungen komplizierte Berechnungen über Wahrscheinlichkeitswerte oder Erwartungswerte anzustellen. Statt dessen greifen Menschen auf Heuristiken zurück – z.B., in dem sie, solange es nicht zu größeren Unregelmäßigkeiten kommt, bei einer einmal gefundenen Einstellung zu einer anderen Person bleiben, oder etwa eher Indizien zur Kenntnis nehmen, die ihre Meinung bestärken als solche, die eine einmal gefundene Meinung falsifizieren könnten. Good führt eine große Zahl psychologischer Studien an, die diese in den Kognitionswissen-

---

[7] Um die Zuspitzung etwas abzumildern: Vielleicht reicht es auch schon aus, dass Agenten sich unterschiedlich freundlich verhalten können, damit Mechanismen, die die soziale Funktion von Vertrauen erfüllen sollen, sinnvoll implementiert werden können.

schaften wohl bekannten Effekte (*set-effect*, *confirmation-bias*) bestätigen, und leitet daraus weitere Bedingungen für erfolgreiche Kooperationen bzw. für Vertrauen ab (vgl. Good 1988: 37ff).

Soziale Einbettung

Zu diesen psychologischen Argumenten kommt die Tatsache, dass Menschen immerzu in soziale und kulturelle Kontexte eingebettet sind. Dies bleibt nicht ohne Auswirkungen auf das menschliche Verhalten; vielfältige soziale und kulturelle Faktoren interagieren miteinander (vgl. dazu Gambetta 1988b: 220ff). Einen einzigen Mechanismus – etwa Vertrauen – aus dem sozialen Gefüge herauszugreifen, und ihn dann vielleicht noch, ungeachtet der vielfältigen Arten von Vertrauen (vgl. Deutsch 1973; Luhmann 1989; Misztal 1996), auf ganz bestimmte Aspekte zu beschränken, birgt deswegen immer die Gefahr, wesentliches zu übersehen. Auch dies gilt es zu berücksichtigen, wenn soziale Mechanismen auf technische Systeme übertragen werden sollen. Ein Beispiel dafür ist die Tatsache, dass Vertrauen in der Luhmann'schen Konzeption auf Hintergrundwissen über die Existenz von »durchschlagskräftiger – um nicht zu sagen: machtvolle[n] – Institutionen« (Bachmann 1998: 223) angewiesen ist. Agenten in einem technischen System verfügen nicht über dieses Hintergrundwissen, und entsprechend vorsichtig und kritisch konzipiert Bachmann (1998) auch seine Vorschläge, wie im Sinne eines vergleichenden Experiments verschiedene Arten von Vertrauen in Multiagenten-Systemen implementiert werden könnten.

Simulation oder …

Soziologische Theoriebildung – und mit ihr auch mathematische oder algorithmische Formalismen und Simulationen – geht zumeist von der Tatsache aus, dass eine Theorie nicht alle Facetten der sozialen Realität abbilden kann, sondern sich auf bestimmte Ebenen oder bestimmte Perspektiven beschränken muss. Eine gute Theorie in diesem Sinne – das gilt leider beileibe nicht für alle soziologischen Theorien – expliziert ihre Grenzen und vergisst nie, dass sie eben immer nur eine komplexitätsreduzierte, zumeist auf idealtypischen Vorstellungen aufbauende Annäherung an die soziale Realität darstellen kann. Mit diesen Gedanken lässt sich dann etwa Gilbert / Troitzsch (1999) oder Malsch (1997) zustimmen, dass Multiagenten-Systeme und ähnliche Ansätze zur Simulation von Gesellschaften durchaus auch einen soziologischen Erkenntnisgewinn mit sich bringen können – ein Gedanke, der in der Soziologie allerdings alles andere als unumstritten ist.

… technische Lösung?

Ein wenig anders stellt sich die Situation dar, wenn nicht soziale Simulationen in den Mittelpunkt der Aufmerksamkeit gerückt werden, sondern der Versuch, die technifizierte Version sozialer Mechanismen zur Lösung (sozio)technischer Probleme heranzuziehen – wenn es also beispielsweise darum geht, wie das Problem der Sicherheit in offenen Multiagenten-Systemen behandelt werden soll. Im Vergleich den simulierten Gesellschaften ändert sich hierbei vor allem ein entscheidender Punkt: Es geht nicht mehr darum, einen sozialen Mechanismus möglichst exakt nachzubilden, sondern der soziale Mechanismus ist hier nur die Anregung, um zu bestimmten Lösungsideen für bestimmte Probleme zu kommen. Die Kehrseite dieses entscheidenden Unterschieds besteht darin, dass wir es

zumeist auch mit noch sehr viel stärker eingeschränkten Gegebenheiten zu tun haben – selbst im Vergleich zum Verhältnis zwischen simulierter und real existierender Gesellschaft – und dass sich so unausweichlich die Frage stellt, ob angesichts dieser Umstände die reduzierte Nachbildung sozialer Mechanismen überhaupt noch zu den gewünschten Ergebnissen führt. Möglicherweise steht hier die 'Angewandte Sozionik' vor ähnlichen Problemen, wie sie sich in der Bionik in Bezug auf die Übernahme von biologischen Gegebenheiten in die Technik stellen. Um es deutlich zu machen: Neben dem bekannten Erfolgen der Bionik – etwa die Feststellung, dass die spezifische Oberflächenstruktur von Haifischen auch bei Booten zu einem besseren Strömungsverhalten führt – gibt es eben auch Mechanismen, die sich nicht so einfach isoliert und komplexitätsreduziert in die Technik übernehmen lassen, beispielweise die Flugbewegung von Vögeln im Vergleich zur Antriebstechnik von Flugzeugen, oder die Photosynthese als Prozess der Energiegewinnung. In Bezug auf sozionische Übernahmen in die Informationstechnik stellt sich hier beispielsweise die Frage, ob die anthropomorphisierende Metapher des *sozialen Agenten* bzw. der *Agenten-Gesellschaft* nicht auch zu überzogenen Erwartungen und Fehlschlüssen auch auf Seite der Programmierenden führt (vgl. auch Malsch et al. 1998; Kuhlen 1999b).

3.3 Schlussfolgerungen in Bezug auf AVALANCHE

Aus diesen Überlegungen ergibt sich in Bezug auf die Ausgangsfrage, ob und wie sich Vertrauen / Vertrauenswürdigkeit / Reputation am besten modellieren lässt, das Problem, dass das hier verwendete Multiagenten-System AVALANCHE einen gewissen Zwitterstatus zwischen Simulation (Modellierung von Wertschöpfungsketten) einerseits und beispielhafter Implementierung oder gar Prototyp eines agentenvermittelten elektronischen Marktplatzes andererseits innehat (vgl. Sackmann 1998; Eymann et al. 1998; Eymann / Padovan 1999). Je nachdem, in welche Richtung dieses Pendel ausschlägt, ergeben sich unterschiedliche Anforderungen an die mögliche Umsetzung sozialer Mechanismen. Im ersten Fall geht es darum, zu simulieren, wie sich betrügerisches Verhalten auf einen Markt auswirkt, und wieweit verschiedene soziale Mechanismen – etwa eine Steuerung von Kooperationen über Vertrauen – die Auswirkungen davon eindämmen können. Im zweiten Fall liegt das Interesse eher darauf, tatsächlichen, letztlich monetären Schaden abzuwenden und eine an soziale Gegebenheiten angelehnte Kontrolle der Stellvertreterhandlungen der Agenten zu erreichen, um so reale Betrügereien um reales Geld zu verhindern.

Marktsimulation

Wenn AVALANCHE primär als Experimentierfeld zur Simulation sozialer Gegebenheiten gesehen wird – beispielsweise unter der Fragestellung, inwieweit »die Variation von Parametern wie Transaktionskosten, Konnektivität oder Markttransparenz […] zu eher marktlichen oder hierarchischen Strukturen führt« (Eymann / Padovan 1999: 626) – dann geht es bei der Hinzufügung einer Simulation von Vertrauensbeziehungen darum, wieweit diese als weitere Komponente in einem ökonomischen Modell Veränderungen beispielsweise in den Kooperations- oder Koordinationsstrukturen hervorrufen. Das Hauptinteresse

würde dann also darin liegen, die – sozialwissenschaftlichen, hier vor allem die ökonomisch relevanten – Merkmale von Vertrauensbeziehungen möglichst zutreffend in die Simulation einzubauen. Damit würde die Komplexität von AVALANCHE etwas vergrößert; weitere, in eine ähnliche Richtung gehende Maßnahmen könnten in der Hinzufügung von Institutionen wie Banken oder in der Einführung einer Marktaufsicht liegen. Der Einbezug von Vertrauensbeziehungen wäre dann eine Erweiterung bzw. Modifizierung der in (Sackmann 1998: 45f) vorgeschlagenen Reputationskoeffizienten. Kriterien zur Evaluation der verschiedenen existierenden Modelle zur Formalisierung von Reputation bzw. Vertrauensbeziehungen wären dann primär die möglichst exakte Abbildung der sozialwissenschaftlichen Vorstellungen über Vertrauensbeziehung bei Kompatibilität zu den AVALANCHE zugrundeliegenden ökonomischen Theorien[8], sowie sekundär eine möglichst gut handhabbare informatische Implementierbarkeit und prinzipiell die daran anschließenden Faktoren wie etwa die Zeitkomplexität des verwendeten Algorithmus.

Marktplatz-Prototyp

AVALANCHE kann allerdings auch als Prototyp für eine im Internet verwendbare, offene Marktplattform für den Handel zwischen Agenten betrachtet werden (vgl. Padovan et al. 2000). Unter diesem Gesichtspunkt würde weniger die wissenschaftliche Exaktheit der Modellierung im Vordergrund stehen, als vielmehr der Aspekt der Sicherheit, und mit der Frage des Vertrauens zwischen den Agenten dann auch die Frage des Vertrauens in den Marktplatz als technischem Gesamtsystem (vgl. u.a. Rasmusson et al. 1997; Kuhlen 1999b). Wichtig wäre dabei die Frage, wieweit soziologische Modelle über die Bildung von Vertrauen oder Reputation auf technische Systeme übertragbar sind, welche Teile des sozialen Mechanismus beispielsweise wegfallen und welche Änderungen in der Funktionalität sich dadurch ergeben. An oberster Stelle steht also das Kriterium, ob die gewünschte Funktion – also ein vertrauenswürdiges, weil sicheres Handelssystem – im sozio-technischen Multiagenten-System durch die ausgewählte Formalisierung erreichbar ist und wo die jeweiligen Schwächen liegen. Die Entsprechung zwischen soziologischer Theorie und technischer Implementierung spielt nur bezüglich der Funktionalität des Multiagenten-Systems eine Rolle, da sie dazu dienen kann, im Abgleich zwischen dem Gültigkeitsbereich der soziologischen Theorie und dem Gültigkeitsbereich der Formalisierung Einschränkungen und Mängel aufzuzeigen. Danach kommen dann wiederum – und hier mit einem höheren Gewicht als bei der Simulation – die Fragen der Implementierbarkeit.

Fazit

Obwohl AVALANCHE als Projekt zur Modellierung wirtschaftlicher Zusammenhänge in der ökonomischen Theorie gestartet ist, scheint es mir gerade auch im Hinblick auf zukünftige Entwicklungen fruchtbarer, im weiteren Verlauf der Arbeit AVALANCHE

---

[8] Als theoretische Grundlage für die Formalisierung von Vertrauen würde sich dann wahrscheinlich eher der Versuch einer *Ökonomik des Vertrauens* (Ripperger 1998) als die hier beschriebenen soziologischen Theorien eignen. Siehe auch S. 34.

Priorität	Markt-Simulation	Marktplatz-Prototyp
1	Entspricht die gewählte Formalisierung der zugrundeliegenden soziologischen Theorie?	Führt die gewählte Formalisierung zu einem sichereren Gesamtsystem? Erfüllt sie die in sie gesetzten Ansprüche?
2	Ist die der gewählten Formalisierung zugrundeliegende Theorie mit den AVALANCHE begründenden ökonomischen Theorien kompatibel?	Sind die von der soziologischen Theorie gemachten Annahmen bei einer Übertragung in das technische System noch gültig?
3	Technische Kriterien (Zeitkomplexität, Berechenbarkeit, Implementierbarkeit) (eher unwichtig)	Technische Kriterien (Zeitkomplexität, Berechenbarkeit, Implementierbarkeit) (wichtig)

Tabelle 2 - Unterschiedliche Kriterien für AVALANCHE als Simulation oder Prototyp

unter dem Gesichtspunkt des Marktplatz-Prototypen zu betrachten. Dementsprechend werden für den Vergleich der verschiedenen Ansätze zur Implementierung von Vertrauen und Reputation im folgenden Kapitel nicht die für soziale Simulationen kennzeichnenden Kriterien herangezogen, sondern die für die Übernahme sozialer Mechanismen in technische Systeme angelegten Maßstäbe. Der Schwerpunkt beim Vergleich der verschiedenen Modelle wird dabei auf den funktionalen Kriterien liegen; die technischen Kriterien können nur am Rande berücksichtigt werden. Mit dieser Entscheidung wird zugleich angenommen, dass die von einem Vertrauens- oder Reputationsmechanismus zu behandelnde Problemlage der eines offenes Systems autonomer und möglicherweise böswilliger Agenten entspricht, wie dies in Abschnitt 3.1 dargestellt wurde. Ein weiterer Effekt dieser Schwerpunktsetzung liegt darin, dass es beim Vergleich der Formalisierungen nicht darum gehen kann, ihren erkenntniserweiternden Gehalt für die Soziologie (vgl. Marsh 1994a: 80f; Gilbert / Troitzsch 1999) zu untersuchen, sondern dass die im Kapitel 2 dargestellte mitgeführte soziologische Theorie den Hintergrund dafür bilden muss, um zu analysieren, ob die verschiedenen Formalismen ihren Ansprüchen an die Etablierung der Funktionalität von Vertrauen in einem technischen System auch gerecht werden können.

4 Formalisierungen von Reputation und Vertrauen

'gedanken papers'

Ziel dieses Kapitels ist es, verschiedene Vorschläge für die Formalisierung von Reputation (bzw. Vertrauenswürdigkeit) darzustellen, miteinander zu vergleichen und anhand der im Kapitel 3 gefundenen Kriterien zu bewerten. Inzwischen gibt es eine umfangreiche Sammlung an Vorschlägen, modellhaften Implementierungen und existierenden Systemen für ein agentenbasiertes Vertrauensmanagement bzw. für Reputationsmechanismen im Bereich von Multi-Agentensystemen, *eCommerce* und Online-Gemeinschaften (Abdul-Rahman / Halles 1997; Bachmann 1998; Foner 1999; Jones / Marsh 1997; Marsh 1994a; Olsson 1998; Rasmusson 1996; Schillo 1999; Winter 1999; Winsborough et al. 2000; Wong / Sycara 1999; Yu et al. 2000; Yu / Singh 2000; Zacharia 1999). Dabei variiert der Grad der Detaillierung von bloßen Ideenskizzen über mathematische Formalisierungen, die

allerdings einige zentrale Implementierungsfragen unbeantwortet lassen bis hin zur Darstellung von Algorithmen in Pseudo-Code. Im Rahmen dieser Arbeit ist es nicht möglich, alle diese Ansätze im Detail darzustellen; insbesondere, da bei einigen der eher in den Bereich der Ideenskizzen – etwas abschätzig könnte bei einigen Ansätzen mit dem *Jargon File* (4.2) teilweise wohl auch von '*gedanken papers*'[9] gesprochen werden – eine detaillierte Darstellung gleichbedeutend mit einer umfangreichen Ausarbeitung der in den Quellen bloß angedeuteten Algorithmen und ihren Umsetzungsmöglichkeiten ist.

Kategorien bei Winter

Um zu einer durchführbaren und trotzdem sinnvollen Darstellung zu kommen, sollen deswegen im folgenden die verschiedenen Ansätze im Rahmen verschiedener größerer Kategorien betrachtet werden. Verschiedene AutorInnen (u.a. Kuhlen 1999b; Rasmusson et al. 1997; Rasmusson / Janson 1996; Winter 1999; Yu / Singh 2000; Zacharia et al. 1999) haben Vorschläge gemacht, wie ein derartiger Rahmen aussehen kann. Brauchbar erscheint dabei vor allem die auf Zacharia et. al (1999) und Rasmusson / Janson (1996) aufbauende Unterteilung nach Winter (1999), wie sie in Abb. 3 dargestellt ist. Unter der Kategorie *'hard' social control* versteht Winter (1999: 142f) in Abgrenzung zu traditionellen harten Mechanismen wie Passwortschutz oder der Verschlüsselung von Daten die im Agentensystem stattfindende soziale Kontrolle, die nicht auf Vertrauen basiert, sondern auf der Institutionalisierungen von Normen, wie etwa einem Rückerstattungs- oder einem Ausschlusssystem.

Die von Kuhlen (1999b: 358ff) vorgeschlagene Unterscheidung von *trust centern* auf der einen Seite und Vertrauensnetzwerken auf der anderen Seite lässt sich als Untermenge in Figur 4 wiederfinden, wobei die *trust center* je nach Ausprägung eher bei den – 'weichen' – zentralisierten Reputationssystemen oder im Bereich der *hard control* anzusiedeln sind. Yu / Singh (2000) legen zum einen einen Schwerpunkt auf die Unterscheidung von Vertrauen als Ergänzung zu *hard security*. Diese Unterscheidung ist in Figur 4 in der Trennung zwischen *soft control* und *hard control* wiederzufinden. Zum anderen betonen sie den Unterschied zwischen *trust networks* und Ansätzen zur sozialen Kontrolle durch Reputation nach Rasmusson / Janson (1996).

---

[9] »gedanken: […] adj. Ungrounded; impractical; not well-thought-out; untried; untested. 'Gedanken' is a German word for 'thought'. A thought experiment is one you carry out in your head. In physics, the term 'gedanken experiment' is used to refer to an experiment that is impractical to carry out, but useful to consider because it can be reasoned about theoretically. […] Gedanken experiments are very useful in physics, but must be used with care. It's too easy to idealize away some important aspect of the real world in constructing the 'apparatus'. […] accordingly, the word has a pejorative connotation. It is typically used of a project, especially one in artificial intelligence research, that is written up in grand detail (typically as a Ph.D. thesis) without ever being implemented to any great extent. Such a project is usually perpetrated by people who aren't very good hackers or find programming distasteful or are just in a hurry. A 'gedanken thesis' is usually marked by an obvious lack of intuition about what is programmable and what is not, and about what does and does not constitute a clear specification of an algorithm. See also AI-complete, DWIM.« (Jargon File 4.2, zit. nach http://www.science.uva.nl/~mes/jargon/g/gedanken.html)

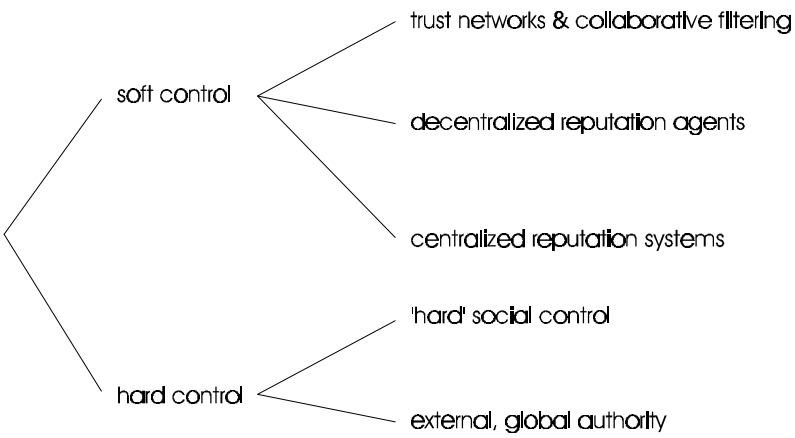


Abb. 3 – Unterteilung der Kontrollmechanismen (nach Winter 1999)

Diskussion

Eignet sich die in Abb. 3 dargestellte Unterteilung als Rahmung für die Analyse der einzelnen Reputations- und Vertrauensmechanismen? Zum einen fällt auf, dass solitäre Ansätze, bei denen Agenten nur aufgrund ihrer eigenen Erfahrungen über die Vertrauenswürdigkeit anderer Agenten entscheiden, nicht so recht in das Raster passen wollen. Zum anderen ist auch bei einer oberflächlichen Sichtung der zu behandelnden Mechanismen eine deutliche Ungleichverteilung auszumachen, etwa was die noch kaum vorhandenen Ansätze für dezentrale Reputationsagenten betrifft. Dem ersten Einwand ist leicht mit einer Erweiterung der Kategorie der ‘subjektiven’ Ansätze – soziale Netzwerke, Vertrauensnetze, etc. – um die solitären Ansätze gegenüberzutreten. Allerdings verstärkt sich das Ungleichgewicht, was die Verteilung der einzelnen Ansätze auf die einzelnen Kategorien betrifft, dadurch noch mehr, was letztlich den Versuch einer Kategorisierung ad absurdum führen würde. Dieses Ungleichgewicht deutet darauf hin, dass im Rahmen dieser Arbeit eine andere Kategorisierung gefunden werden muss, die die Skala von lokalen ‘subjektiven’ bis hin zu globalen, ‘objektiven’ Ansätze durch andere Dimensionen ergänzt oder ersetzt.

Alternativvorschlag

Aus der Perspektive eines einzelnen Agenten wäre so zwischen Ansätzen zu unterscheiden, die in einem unterschiedlich hohen Maß die Arbeit zur Ermittlung der Reputation auf Außenstehende verlagern. Unterschieden werden könnte dann zwischen der überwiegend durch den Agenten selbst zu leistenden Informationsarbeit einerseits und dem überwiegenden Rückgriff auf externe Agenturen andererseits. Aus objektorientierter Sicht würde dieser Unterscheidung danach fragen, ob die Berechnung der Vertrauenswürdigkeit eine Methode innerhalb der allgemeinen Agentenklasse ist, oder ob diese Methode auf andere Objekte (Handelsplätze, spezialisierte Agenten, Gesamtsystem) ausgelagert wurde. Eine quer dazu liegende Dimension könnte berücksichtigen, welche Rolle die Bewertungen dritter für die Berechnung der Reputation oder der Vertrauenswürdigkeit haben. Dabei ergibt sich dann folgende Kreuztabelle (Tabelle 3).

	Reputationsberechnung überwiegend intern	Reputationsberechnung überwiegend extern
Bedeutung der Bewertungen dritter ist gering	(1) agentenzentrierte solitäre Ansätze	(3) 'objektive' externe Bewertungsagenturen
Bedeutung der Bewertungen dritter ist groß	(2) agentenzentrierte soziale Ansätze	(4) 'subjektive' externe Bewertungsagenturen

Tabelle 3 – Ein alternativer Kategorisierungsversuch

Im folgenden werden nun Ansätze aus den vier Feldern dieser Kreuztabelle diskutiert. Der Schwerpunkt liegt dabei auf den agentenzentrierten Ansätzen. Dabei werde ich jeweils eine Formalisierung ausführlicher darstellen, und Alternativen dazu im selben Feld kurz diskutieren. Dabei sollte mitgedacht werden, dass die hier getroffenen Unterscheidungen vorhandene, fließenden Übergänge zwischen den einzelnen Kategorien weitgehend ignorieren, da sie das Ziel haben, eine analytische Hilfestellung bei der Betrachtung der verschiedenen Vorschläge zu geben.

4.1 Agentenzentrierte, solitäre Ansätze

Abgrenzung

Agentenzentrierte, solitäre Ansätze zeichnen sich dadurch aus, dass die Berechnung der Vertrauenswürdigkeit potenzieller Kooperationspartner zum einen vom Agenten selbst durchgeführt und bewertet wird, und dass zum anderen nur auf Erfahrungen aus eigenen Begegnungen zurückgegriffen wird. Das Kernstück des derartigen Ansätzen zugrundeliegenden Algorithmus lässt sich aus Sicht des berechnenden Agenten kurz wie in Tabelle 4

```
procedure Reagiere-auf-Kooperationsangebot(agent, situation, importance, utility);
begin
    if memory(agent) = NIL then
        füge ein Modell für agent in memory ein
        memory[agent]. vertrauenswürdigkeit[0] ← Defaultwert
    end
    v ← Abschätzung der Vertrauenswürdigkeit ausgehend von den in memory
          gespeicherten zeitbezogenen generellen Vertrauenswerten, sofern die
          Situation zu diesem Zeitpunkt der jetzigen ähnelte
    sv ← Berechne-situative-Vertrauenswürdigkeit(v, situation, importance, utility)
    ks ← Berechne-Kooperationsschwelle(v, situation, importance)
    if sv > ks then
        Interagiere mit dem Agenten
        v ← memory(agent).vertrauenswürdigkeit[t]
        Erhöhe v um Vertrauensfaktor
        memory(agent).vertrauenswürdigkeit[t+1] = v
    else
        v ← memory(agent).vertrauenswürdigkeit[t]
        Reduziere v um Misstrauensfaktor
        memory(agent).vertrauenswürdigkeit[t+1] = v
    end
    Speichere Informationen über die Situation zum Zeitpunkt t+1
end
```

Tabelle 4 – Kernstück eines agentenzentrierten, solitären Ansatzes

gezeigt beschreiben. Als Reaktion auf Anfragen anderer Agenten wird die Vertrauenswürdigkeit des anderen Agenten mit weiteren Entscheidungsfaktoren, etwa dem möglichen Nutzen und der geschätzten Wichtigkeit der Handlung zu einer (evtl. situationsabhängigen) Vertrauenswürdigkeit verrechnet und mit einer Vertrauensschwelle verglichen. Kommt es zu einer Handlung mit dem anderen Agenten, wird dessen Vertrauenswürdigkeit je nach Ausgang der Handlung angepasst; je nach Ansatz wird dabei bereits das Ablehnen eines Kooperationsangebots mit Vertrauensverlust bestraft. Ähnlich sieht auch der Algorithmus für die Entscheidung aus, welchem von mehreren möglichen Kooperationspartnern ein Kooperationsangebot gemacht wird – hier wird der Agent ausgewählt, dessen (evtl. situationsspezifische) Vertrauenseinschätzung (evtl. bezogen auf einen situations- und agentenabhängigen Schwellenwert) am besten ist, und diesem dann ein Kooperationsangebot gemacht. Je nach Zustimmung oder Ablehnung kann dann wiederum das gespeicherte generelle Vertrauen in diesen Agenten angepasst werden.

Alternativ können die gemachten Erfahrungen auch in ein über die Speicherung der zeitpunktspezifischen Vertrauenswürdigkeit hinausgehendes Modell des anderen Agenten eingefügt werden. Daneben spielt es für diesen Algorithmus eine große Rolle, wieweit das Gedächtnis zurückreichen soll, bzw. wie schnell gemachte Erfahrungen vergessen werden sollen. Ein weiterer wichtiger Punkt bezieht sich auf bisher unbekannte Agenten. Da hier keine Erfahrungen und somit weder ein Wert für deren Vertrauenswürdigkeit und erst recht kein Modell über ihr Verhalten existiert, lassen sich auf der Basis eigener Erfahrungen erst einmal keine Aussagen machen. Deswegen muss in einem solchen Fall eine Entscheidung darüber getroffen werden, wie groß die Vertrauenswürdigkeit von Unbekannten ist bzw. welche Heuristiken herangezogen werden sollen.

4.1.1 Stephen Marsh – Trust mit großem T

Marsh 1994a

Wohl der bekannteste in diese Kategorie fallende Ansatz ist die von Stephen Marsh (1992, 1994a, 1994b; Jones / Marsh 1997) vorgeschlagene Formalisierung von *Trust*[10]. Insbesondere in *Formalising Trust as a Computational Concept* stellt Marsh (1994a) sein mathematisches Konzept – ausgehend von soziologischen und psychologischen Vertrauenskonzepten – ausführlich dar. Die Funktionsweise des Formalismus wird anhand einiger Fallbeispiele sowie einer Implementierung für modifizierte Gefangenendilemma-Spiele in einer sehr einfachen Agentengesellschaft demonstriert. Marshs Konzeption soll hier nun ausführlicher dargestellt werden.

---

[10] Diese Bezeichnung (*Trust* im Gegensatz zu *trust*) habe ich aus (Jones / Marsh 1997: Introduction) übernommen: »We call our formal description Trust in order to differentiate it from wider definitions.«; Marsh selbst (1994a) trifft diese Unterscheidung nicht, sondern geht von einer allgemeingültigen Formalisierung von Vertrauen aus..

Situatives Vertrauen

Kernstück der Vertrauensberechnung ist folgende Formel, mit der das situative Vertrauen $T_x(y, \alpha)$ des Agenten *x* in den Agenten *y* in der Situation[11] α aus dem subjektiven Nutzen $U(\alpha)$ – im Sinne eines Erwartungswertes für den Nutzen der Situation α – und der subjektiven Bedeutung der Situation $I(\alpha)$ berechnet wird (Marsh 1994a: 62):

$$T_x(y,\alpha) = U_x(\alpha) \times I_x(\alpha) \times \hat{T}_x(y)$$

Der Nutzen $U(\alpha)$ kann dabei Werte aus dem Intervall [-1,+1] annehmen, während die Wichtigkeit oder Bedeutung der Situation $I(\alpha)$ Werte aus dem Intervall [0,1] annehmen kann. Der letzte Ausdruck der Formel, $\hat{T}_x(y)$, bezieht sich auf einen möglicherweise gewichteten Mittelwert über die Bewertungen der generellen Vertrauenswürdigkeit des Agenten *y* durch den Agenten *x* in der Vergangenheit ($T_x(y)^i$, …, $T_x(y)^{t-1}$). Er kann – wie die Vertrauensfunktion selbst auch – Werte aus dem Intervall [-1,+1)[12] annehmen. Marsh diskutiert zur Berechnung dieses Ausdrucks verschiedene 'Dispositionen' – optimistische, pessimistische oder realistische Agenten (1994a: 65ff; 1994b). In allen Fällen geht es dabei darum, wie der Agent die Vertrauenseinschätzungen in der Vergangenheit (bis zu einem gewissen, durch seine Gedächtnisspanne oder durch die Zahl bisheriger Begegnungen eingeschränkten Zeitpunkt) bewertet. So würde ein optimistischer Agent den Maximalwert aller bisherigen Begegnungen in ähnlichen Situation als Grundlage für die Schätzung nehmen, ein pessimistischer Agent den Minimalwert. Als Grundlage für einen realistischen bzw. pragmatischen Agenten zieht Marsh den Mittelwert über die situative Bewertung dieses Agenten in ähnlichen Situationen in der Vergangenheit heran.

Vertrauensschwelle

Um den Schwellenwert zu berechnen, ab dem ein situativer Vertrauenswert als ausreichend groß angesehen wird, um tatsächlich zu vertrauen ($T_x(y, \alpha) > Cooperation\_Treshold_x(\alpha)$), verwendet Marsh (1994a: 69) folgende Gleichung:

$$\text{Cooperation\_Treshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha) + \hat{T}_x(y)} \times I_x(\alpha)$$

Dabei spielt sowohl das wahrgenommene Risiko einer Situation α ($Perceived\_Risk_x(\alpha)$) als auch die Kompetenz, die der Agent *x* dem Agenten *y* in der Situation α zuschreibt, eine wichtige Rolle. Für das wahrgenommene Risiko kommt Marsh (1994a: 71f) zu einer recht umfangreichen Abwägung je nach Wissenstand des Agenten über die Situation (kein / teilweise vorhandenes / komplettes Wissen). So soll ein Agent beispielsweise, wenn

---

[11] Mit Situation (Marsh 1994a) bzw. Kontext (Jones / Marsh 1997) ist die Vergleichbarkeit von Handlungen gemeint. Vertrauenswürdigkeit bezieht sich immer auf die Erfahrungen in vergleichbaren Situationen. Bezogen auf einen elektronischen Marktplatz könnte dies z.B. heißen, dass Vertrauenswürdigkeit bezüglich Kaufsituationen unabhängig von der Vertrauenswürdigkeit bezüglich Verkaufsituationen ist. Darüber, ob diese Annahme absolut genommen realistisch ist, lässt sich diskutieren.

[12] Marsh schließt den Wert +1 aus dem Intervall aus, da dieser blindes Vertrauen symbolisieren würde, was er – unter Bezugnahme u.a. auf Luhmann – nicht für sinnvoll hält (vgl. Marsh 1994a: 57f).

überhaupt kein Wissen vorhanden ist, auf seine eigene Risikoeinschätzung bezüglich früherer Situationen ohne Wissen zurückgreifen, und so bei einer Fehleinschätzung zumindest sein Risikobewusstsein stärken. Soweit ich sehe, verwendet Marsh in der Praxis allerdings keine berechneten, sondern gesetzten Werte für *Perceived_Risk*. Ähnlich sieht es mit der *Perceived_Competence* aus, für die Marsh (1994a: 73f) ebenfalls drei Fälle unterscheidet[13], die zwar ausführlich diskutiert werden, in seinen eigenen Beispielen aber doch eher durch Abschätzungen als durch Berechnungen bestimmt werden.

Auswahl des Partners

Um einen Agenten für eine Kooperation auszuwählen, wenn mehrere Agenten die Vertrauensschwelle überschreiten, diskutiert Marsh (1994a: 91) verschiedene Verfahren. So kann der generell vertrauenswürdigste Agent gewählt werden oder auch der in dieser Situation vertrauenswürdigste. Es wäre auch möglich, Vertrauen ganz außer Spiel zu lassen und den Agenten mit dem niedrigsten *Cooperation_Treshold* auszuwählen. Ein weiteres Verfahren könnte darin bestehen, den Agenten mit der maximalen Distanz zwischen *Cooperation_Treshold* und situativem Vertrauen auszuwählen, oder gar eine Entscheidung davon abhängig zu machen, dass ein Agent in möglichst vielen der genannten Verfahren am besten abschneidet.

Unbekannte Agenten

Steht ein Agent bei Marsh einem ihm bisher unbekannten Agenten gegenüber, ändert sich sowohl die Berechnung des situativen Vertrauen wie auch die des Schwellenwerts für Kooperationen. Was bei Marsh (1994a) eher implizit vorausgesetzt wird, wird in (Jones / Marsh 1997) expliziert. Für einen völlig unbekannten Agenten entspricht die Abschätzung der Vertrauenswürdigkeit $\hat{T}_x(y)$ dem *basic trust* T_x des Agenten *x*, also seiner 'Disposition', Vertrauen auszusprechen. Ist der Agent bekannt, aber nicht in dieser Situation, so kann die Abschätzung der Vertrauenswürdigkeit durch das generelle Vertrauen in den anderen Agenten ersetzt werden. Ähnlich sieht es bei der Formel für den *Cooperation_Treshold* aus, wo sowohl die *Perceived_Competence* (vgl. Fußnote 13) als wieder-

---

[13] Wenn der Agent *x* den Agenten *y* nicht kennt, schätzt er dessen Kompetenz mit folgender Formel ab (T_x bezieht sich dabei auf den sogenannten *basic trust* des Agenten, also seine allgemeine situations- und partnerunabhängige Neigung zu Vertrauen):

$$\text{Perceived\_Competence}_x(y,\alpha) = T_x I_x(\alpha)$$

Ist dagegen der Agent *y* schon bekannt, kann auf die bereits gemachten Erfahrungen zurückgegriffen werden, die *x* mit *y* entweder in unähnlichen Situationen gemacht hat –

$$\text{Perceived\_Competence}_x(y,\alpha) = \frac{1}{|A|}\sum_{\beta \in B}(\text{Experienced\_Competence}_x(y,\beta)^{t'}) \times \hat{T}_x(y)$$

oder sogar auf die dann nicht mit Vertrauensabschätzungen zu verrechnenden Erfahrungen in ähnlichen Situationen:

$$\text{Perceived\_Competence}_x(y,\alpha) = \frac{1}{|A|}\sum_{\alpha \in A}(\text{Experienced\_Competence}_x(y,\alpha)^{t'})$$

um $\hat{T}_x(y)$ angepasst werden. Für den Fall eines völlig unbekannten Agenten *y* ergibt sich dann folgende nur von der Situation und dem *basic trust* abhängige Ungleichung:

$$U_x(\alpha) \times I_x(\alpha) \times T_x > \frac{\text{Perceived\_Risk}_x(\alpha)}{T_x + T_x / I_x(\alpha)} \Rightarrow vertraue(y)$$

Optimismus/Pessimismus

Wie schon erwähnt, unterscheidet Marsh (1994a: 65ff; 1994b) zwischen verschiedenen 'Dispositionen', die sowohl Einfluss auf den *basic trust* eines Agenten und dessen weitere Entwicklung haben (vgl. Marsh 1994a: 56), als auch im Hinblick auf die Berechnung der Abschätzung der vergangenen Vertrauenserfahrung zur Auswahl unterschiedlicher Mechanismen führen. In seiner Beispielanwendung (*Iterated Prisoner Dilemma*) lässt Marsh Agenten mit verschiedenen Dispositionen gegeneinander antreten (1994a: 110). Weiterhin diskutiert er die Auswirkungen, die die verschiedenen Dispositionen auf die Selbstaufrechterhaltung von Vertrauen bzw. Misstrauen haben (Marsh 1994b; 1994a: 97f). Ein Vergleich mit Luhmanns Diskussion der Vertrauens- bzw. Misstrauensverstärkung liegt nahe.

Vertrauensanpassung

Die Veränderung der generellen Vertrauenswürdigkeit ($T_x(y)^t$) nach erfolgter oder misslungener Kooperation macht Marsh (1994a: 78ff) u.a. davon abhängig, wieweit sich die Agenten an frühere Kooperationen oder Nicht-Kooperationen erinnern[14], und natürlich davon, ob es zu erfolgreicher Zusammenarbeit kam oder nicht. Für seine Beispielanwendung (*Iterated Prisoner Dilemma*) verwendet Marsh die in Tabelle 5 angegebenen Veränderungen (aus der Perspektive von Agent *A*), die wohl allerdings noch mit absoluten Kappungsgrenzen bei -1 und +1 versehen werden müssen:

	B kooperiert	B kooperiert nicht
A kooperiert	$T_A{}^{t+1} = T_A{}^t$ x 1,01 $T_A(B)^{t+1} = T_A(B)^t$ x 1,10	$T_A{}^{t+1} = T_A{}^t$ x 0,99 $T_A(B)^{t+1} = T_A(B)^t$ x 0,90
A kooperiert nicht	$T_A{}^{t+1} = T_A{}^t$ x 1,05 $T_A(B)^{t+1} = T_A(B)^t$ x 1,01	$T_A{}^{t+1} = T_A{}^t$ x 0,95 $T_A(B)^{t+1} = T_A(B)^t$ x 0,90

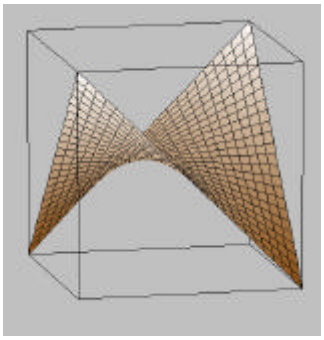
Tabelle 5 – Veränderung von As *basic trust* und As generellen Vertrauens in B

Diskussion

Einer der wichtigsten Einwände gegen diese Modellierung von Vertrauen besteht darin, dass es an mehreren Stellen zu 'seltsamen Effekten' kommt, die der zugrundeliegenden Mathematik geschuldet sind (vgl. Schillo 1999: 26ff). Marsh sieht diese Problemfälle zumindest teilweise durchaus selbst, hält aber daran fest, dass seine Formalisierung Vertrauen erfolgreich modelliert (Marsh 1994a: 143; vgl. auch die Diskussion zu *discrete positive values* in Jones / Marsh 1997). Beispielhaft lässt sich dies an Marshs Formel für

[14] Dies hat insbesondere Folgen, weil Vertrauen bei Marsh gerade nicht als Reaktion auf zuvor entgegengebrachtes Vertrauen verstanden werden soll.

das situative Vertrauen zeigen[15]. Hierbei handelt es sich wie gesehen um ein Produkt, bei dem verschiedene Faktoren negativ werden können. Dies hat den etwas seltsamen Nebeneffekt, dass das situative Vertrauen positiv wird, wenn sowohl der Nutzen der Situation negativ gesehen wird als auch dem anderen Agenten Misstrauen entgegengebracht wird (vgl. Tabelle 6). Marsh (1994a: 63) beschreibt diesen Effekt als *machievellian*, und meint damit, dass es in einem gewissen Sinn nützlich sein kann, eine Handlung, von der kein Nutzen zu erwarten ist, einem Agenten zu überlassen, dem nicht vertraut wird. Damit engt er aber prinzipiell den Gültigkeitsbereich seiner Formel für situatives Vertrauen auf machiavellistisch ausgerichtete Agenten ein. Weitere, ähnlich fragwürdige Effekte der Verwendung von Multiplikation und des Intervalls [-1,+1) bestehen darin, dass Indifferenz bezüglich des Vertrauens ($T_x(y)=0$) oder des Nutzens ($U(\alpha)=0$) das situative Vertrauen auf den schwer zu interpretierenden Wert 0 setzt[16].



$T_x(y,\alpha)$	$U(\alpha)<0$	$U(\alpha)=0$	$U(\alpha)>0$
$T_x(y)<0$	positiv	0	negativ
$T_x(y)=0$	0	0	0
$T_x(y)>0$	negativ	0	positiv

Tabelle 6 – Mögliche Werte der Vertrauensfunktion (nach Marsh 1994a: 63)

Sicheres Gesamtsystem?

Entsprechend den am Ende des dritten Kapitels aufgestellten Kriterien ist es uns allerdings wichtiger, zu klären, ob die von Marsh vorgeschlagene Formalisierung zu einem sicheren Gesamtsystem führt, wenn sie etwa in einem agenten-basierten elektronischen Marktplatz eingesetzt würde, als die Frage zu beantworten, ob sie eine valide Modellierung menschlichen Vertrauens darstellt, oder sogar einen Beitrag zur theoretischen Fundierung der Soziologie liefern kann[17]. Dazu erscheint es sinnvoll, Marshs Modell auf seine Funktions-

---

[15] Ähnliche Argumente lassen sich gegen die Formel für den *Cooperate_Treshold* anbringen, etwa wenn *Perceived_Competence* + *T*(y) = 0 ist oder wenn es zu nicht sinnvoll interpretierbaren Ergebnissen kommt (vgl. Marsh 1994a: 69f).

[16] Vgl. dazu die Ausführungen zu *No Trust and Distrust* bei Marsh (1994a: 56f), wo alleine vier mögliche Bedeutungen der Null – bezogen auf generelles Vertrauen – angeführt werden.

[17] Marsh selbst sieht seine Formalisierung nicht nur als brauchbaren Ausgangspunkt für eine Implementierung in verschiedenen Agentensystemen und im Bereich des *Computer Supported Cooperative Work* (vgl. auch Jones / Marsh 1997), sondern auch als Beitrag, um innerhalb von *Distributed Artificial Intelligence (DAI)*, Sozialpsychologie und Soziologie zu einem genauer definierten Modell von Vertrauen zu kommen (Marsh 1994a: 81, 144). Trotz der umfangreichen soziologischen Vorarbeiten, die er in seiner Arbeit vornimmt, ist von einer (positiven) Rezeption der vorgeschlagenen Formalisierung durch die Soziologie wenig zu merken. Selbst in der *DAI*-nahen Sozionik wird sein Anspruch, einen Beitrag zur theoretischen Fundierung von Vertrauen geleistet zu haben, rundweg abgelehnt (»Allerdings kann man auch hier sehen, daß man ohne ernsthaften Bezug auf die soziologische Grundlagentheorie das Phänomen nur erahnen kann.« (Bachmann 1999: 203)).

fähigkeit hin abzuklopfen. Hierbei geht es insbesondere um den Aspekt der sozialen Kontrolle (vgl. etwa Dasgupta 1988b), also beispielsweise darum, inwieweit betrügerische Agenten erfolgreich aus Kooperationen ausgeschlossen werden. (vgl. Marsh 1994a: 138). Prinzipiell scheint die hier diskutierte Formalisierung zumindest dazu geeignet sein, dass ein Agent *x* nach einer betrügerischen Interaktion mit einem Agenten *y* – etwa in der Form, dass dieser, da der Vertrauensalgorithmus offen liegt, erst einmal versucht, sich einzuschmeicheln, um dann, wenn entsprechend großes Vertrauen vorhanden ist, zu betrügen – generell Abstand davon nimmt, wieder mit *y* zu kooperieren. Dazu müssten allerdings die Werte für die Veränderung des generellen Vertrauens angepasst werden, etwa durch eine besonders drastische Reduzierung der generellen Vertrauenswürdigkeit bei Betrugsfällen. Damit dürften zumindest pessimistische und realistische Agenten sich einigermaßen gegen Betrug schützen können; optimistische Agenten würden eher auf zurückliegende positive Bewertungen als auf den aktuellen Betrugsvorfall schauen; je nach Größe des Gedächtnisses kann dies zu einem Problem werden oder auch nicht. Allerdings könnte, da eine Kommunikation von Agenten über ihre Erfahrungen nicht existiert, ein gewiefter Betrüger nach und nach jeden anderen Agenten auf diese Art und Weise zumindest einmal betrügen. Eine Lösungsmöglichkeit dafür gibt es in solitären Ansätzen nicht. Insofern trägt der von Marsh gewählte Ansatz also nur in einem sehr beschränkten Maß zu einem sichereren Gesamtsystem bei. Der Fairness halber muss allerdings erwähnt werden, dass dies auch nicht der Ausgangspunkt von Marshs Überlegungen ist, sondern dass ihm vor allem an der Bildung von Vertrauensbeziehungen und ihrer Auswirkungen auf Kooperationen als an der Sicherheit durch Vertrauen / Misstrauen liegt, und dass er die Lösungsmöglichkeit der Kommunikation zwischen Agenten selbst vorschlägt (vgl. Marsh 1994a: 99).

Gültigkeit der Annahmen

Neben der Frage nach der Funktionalität des Systems stellt sich die Frage, ob die zugrunde gelegten soziologischen Annahmen unter den Einschränkungen eines technischen Systems noch Bestand haben. Als Ausgangspunkt für die Definition von Vertrauen dienen bei Marsh (vgl. 1994a: 25ff) die Arbeiten von Morton Deutsch, Niklas Luhmann, Bernard Barber und Diego Gambetta, wobei besonders der Einfluss des Letztgenannten deutlich ist, was sich etwa in der Annahme eines situations- und agentenspezifischen Schwellenwertes für Kooperation niederschlägt, aber auch in der Konzeption der situativen Vertrauens als Produkt von *Utility* und *Importance* auf der einen Seite und Vertrauenswürdigkeit auf der anderen Seite. Dabei fällt positiv auf, dass Marsh durch den Einbezug der Situation und ihrer Bedeutung in gewisser Weise auf das Argument bei Gambetta (1988b: 220) eingeht, dass soziale Arrangements die Notwendigkeit von Vertrauen stark beeinflussen können. Auch die Veränderungen des Vertrauens nach geglückter bzw. misslungener Zusammenarbeit (vgl.Tabelle 5) entspricht der Annahme der soziologischen Theorie, dass Vertrauen langsam wächst und beim Vorliegen negativer Evidenz schnell fällt. Eine aus Sicht der zugrundeliegenden Theorien möglicherweise problematische Annahme ist die Beschreibung von Vertrauenswürdigkeit durch ein Intervall [-1;+1). Gambetta (1988b) beispielsweise

schlägt das Intervall [0;1] vor. Von anderen AutorInnen ist anzunehmen, dass sie mit einer quantitativen Formalisierung der Vertrauenswürdigkeit eher nicht glücklich geworden wären (vgl. auch Schillo 1999: 26). Marshs Gleichungen implizieren, dass es so etwas wie ein um 10% wachsendes Vertrauen gibt – und ignoriert dabei etwa Schwellenwerte, an denen es zu qualitativen Veränderungen des Vertrauens kommt (vgl. Luhmann 1989: 45ff). Ein Teil der Kritik an dieser Skalierung kann dadurch zurückgewiesen werden, dass Marsh mit dem *Cooperation_Treshold* zugleich einen individuellen und situationsabhängigen Schwellenwert in sein Kalkül einbezieht, und dass er sich viele Gedanken über die Behandlung der Vertrauensveränderung in Bezug auf bisherige Beziehungen macht. Dennoch bleibt ein gewisses Unbehagen bezüglich dieser – mit vielen anderen Ansätzen geteilten – sehr quantitativen Behandlung von Vertrauen bestehen (vgl. dazu auch Marsh 1994a: 142).

Technische Kriterien

Aus technischer Sicht stellt Marshs Ansatz ein recht leicht implementierbares Modell dar, auch wenn wichtige Details in der Formalisierung in (Marsh 1994a) nicht enthalten sind, und je nach gewünschter Implementierung erarbeitet werden müssen, und wenn nicht klar ist, ob konkrete Implementierungen alle Details des Ansatzes berücksichtigen werden. Die Abschätzung der Vertrauenswürdigkeit anderer Agenten ist recht effizient gelöst. Wenn eine Kooperationsentscheidung für einen von n Agenten getroffen werden muss, und maximal m zurückliegende Vertrauensbewertungen erinnert werden, dann liegt die Speicherkomplexität im Bereich von $O(n\ m)$, und die Zeitkomplexität – unter der Annahme, dass Operationen wie der Vergleich einer aktuellen Situation mit einer gespeicherten Situation in linearer Zeit möglich sind – für die Auswahl eines vertrauenswürdigen Partners ebenfalls bei $O(n\ m)$[18].

Fazit

Trust stellt einen relativ einfach zu implementierenden und zu berechnenden Ansatz dar, mit dem Agenten unabhängig von den Erfahrungen anderer Agenten zurückgreifend auf ihre eignen bisherigen Vertrauenseinschätzungen zu Bewertungen der Vertrauenswürdigkeit anderer Agenten kommen können. Je nach Einstellung verschiedener Parameter (z.B. Größe des Verlusts von Vertrauenswürdigkeit bei Betrug, Gedächtnisspanne, Disposition der Agenten) sind die Ergebnisse der Vertrauensbewertung mehr oder weniger realistisch und praktikabel. Problematisch erscheinen Details der gewählten Formalisierungen (etwa das Problem der Multiplikation negativer Werte). Die Entscheidung, Vertrauen alleine von den eigenen Erfahrungen abhängig zu machen, trägt zur einfachen Implementierbarkeit bei, stellt aber zugleich eines der größten Probleme dieses Ansatzes dar.

---

[18] Die Berechnung der *Experienced_Competence* sowie die abgeschätzten generellen Vertrauenswerte für jeden zur Auswahl stehenden Agenten (um *Cooperation_Treshold* und situatives Vertrauen ermitteln zu können) ist sicher in $O(n\ m)$ machbar, die Suche nach dem besten Agenten anhand der Ergebnisse dieser Berechnungen in $O(n)$, so dass insgesamt eine möglicherweise noch weiter optimierbare Zeitkomplexität von $O(n\ m)$ für die Auswahl des besten unter n möglichen Agenten gegeben ist.

4.1.2 Weitere agentenzentrierte, solitäre Ansätze

AVALANCHE

Die bisherigen Ansätze, Vertrauen bzw. Reputation in AVALANCHE einzubauen (vgl. Sackmann 1998; Padovan et al. 2000) geht von einem Modell aus, das einen Reputationskoeffizienten – also Abschätzungen der Reputation, die hier als Wahrscheinlichkeit vertrauenswürdigen Verhaltens gesehen wird – in die im Modell verwendeten Transaktionskosten einbezieht. Padovan et al. (2000: 10) diskutieren eine Erweiterung dieses Ansatzes um die Möglichkeit, domänenspezifische Reputationsinformationen von zentralen Agenturen zu erhalten. Eine genauere Darstellung der von AVALANCHE verwendeten Ansätze erfolgt im Kapitel 5.

Bachmann 1998

Ein weiterer agentenzentrierter, solitärer Ansatz wird – trotz der von ihm vorgebrachten Einwände gegen eine formale Umsetzung der Luhmann'schen – von Reinhard Bachmann (1998) vorgeschlagen. Ziel soll dabei die Modellierung inter-personalen Vertrauens für soziale Simulationen im Rahmen der *Distributed Artificial Intelligence* sein. Dazu führt Bachmann Regeln auf (vgl. Tabelle 7), nach denen sich Agenten verhalten sollen. [19]

1. Gibt stets eine Selbstbeschreibung deines Handelns, in der du deine Motive offenlegst.
2. Formuliere positive Erwartungen in bezug auf das zukünftige Verhalten deines Gegenübers.
3. Formuliere Erwartungen über die Erwartungen, die dein Gegenüber in bezug auf dich selbst, entwickelt.
4. Überprüfe die Häufigkeit deiner richtigen und deiner unrichtigen Erwartungen, indem du das Verhalten deines Interaktionspartners in gewissen zeitlichen Abständen auswertest.
5. Wenn sich eine Tendenz erkennen läßt, daß die Richtigkeit deiner gemachten Erwartungen zunimmt, dann indiziere dein Gegenüber mit einem Wert für X (='Vertrauenswürdigkeit'). Dieser Wert soll nach einer bestimmten Anzahl weiterer enttäuschungsfreier Interaktionen graduell erhöht werden.
6. Umgekehrt: Jede Interaktion, die mit einer Erwartungsenttäuschung endet, soll je nach der vorhergehenden Anzahl von enttäuschungsfreien Interaktionen mit dem jeweiligen Agenten mit einem Wert für Y (='Vertrauensunwürdigkeit') belegt werden. *[Nach Regel 5 und 7 wäre es logischer, hier von einem Wert pro Agent und nicht pro Interaktion auszugehen, T.W.]*
7. Verrechne den Vertrauenswürdigkeitswert mit dem Vertrauensunwürdigkeitswert des jeweiligen Agenten nach einem näher zu spezifizierenden Algorithmus, der nach einer besonders hohen Anzahl von enttäuschungsfreien Interaktionen einen 'no claim bonus' vergibt, also einen Enttäuschungsfall unbewertet durchgehen läßt, und bei Enttäuschungen, die in kürzeren Zeitabständen erfolgt sind, den maximalen Vertrauensunwürdigkeitswert (bzw. minimalen Vertrauenswürdigkeitswert) einsetzt.
8. Kooperiere stets zuerst mit demjenigen unter den für deine Zweck [!] in Frage kommenden Mitagenten, denen du den höchsten Wert für 'Vertrauenswürdigkeit' (bzw. niedrigsten Wert für 'Vertrauensunwürdigkeit') zugerechnet hast.

[19] Ein weiterer, allerdings noch vager als dieser skizzierter Ansatz von Bachmann (1999) zur Simulation institutionell basierten Vertrauens bleibt hier unberücksichtigt, da es dabei vor allem darum geht, die Veränderung institutioneller Regeln in einer Gesellschaft zu simulieren – in Bezug auf ein möglichst sicheres Multiagenten-Marktsystem wohl eher abschreckend!

9. Wäge das Risiko betrogen zu werden mit dem möglichen Nutzen ab, der für dich mit einer enttäuschungsfreien Interaktion verbunden sein kann, soweit du das im Voraus abschätzen kannst.

Tabelle 7 – Regeln für inter-personales Vertrauen (Bachmann 1998: 226f)

Da dieser Proto-Algorithmus sich zum einen (wie im übrigen auch Marsh) auf Thimbleby et al. (1994) beruft, und zum anderen in vielen Punkten sehr skizzenhaft bleibt, erübrigt sich hier eine weitergehende Auseinandersetzung. Einige interessante Ideen sollen allerdings dennoch hervorgehoben werden. So schlägt Bachmann im Gegensatz zu Marsh zwei unterschiedliche Variablen für Vertrauenswürdigkeit und Vertrauensunwürdigkeit vor (Regeln 5, 6) – ob hiermit eine Distinktion zwischen Misstrauensbildung und Vertrauensbildung vorgenommen werden soll, oder ob er sich davon eine bessere Implementierbarkeit verspricht, bleibt leider unklar. Die Idee (Regel 7), nach einer längeren Zeit vertrauensvoller Zusammenarbeit einen Bonus einzuführen, der eine Enttäuschung unberücksichtigt lässt, ist ebenso wie ihre negative Entsprechung ein Schritt hin zu einer weniger linearen Formalisierung von Vertrauen, da damit bestimmte qualitative Schwellen simuliert werden. Auch der Vorschlag, Erwartungen an das Verhalten anderer Agenten im Algorithmus zu explizieren und Selbstbeschreibungen auszutauschen – beides liegt in Luhmanns theoretischen Vorstellungen begründet– verdient eine nähere Betrachtung.

Ripperger 1998

Erwähnt werden soll an dieser Stelle noch die *Ökonomik des Vertrauens* von Tanja Ripperger (1998). Sie entwickelt im Rahmen der ökonomischen Theorie (Transaktionskostenanalyse, Principal-Agent-Theorie) eine allgemeine Formalisierung von Vertrauen als einem »Mechanismus zur Stabilisierung unsicherer Erwartungen und zur Verringerung von Handlungskomplexität« (Ripperger 1998: 36) aus Sicht des Vertrauensnehmers und des Vertrauensgebers ein. Der Ansatz von Ripperger ist nicht agenten-spezifisch im Sinne der Agententheorie der Informatik, sondern modelliert Vertrauensbeziehungen als Erwartungswerte von Personen. Ich führe ihn dennoch an dieser Stelle auf, weil Ripperger den *Prozess* der Vertrauensbildung und -abschätzung aus ökonomischer Sicht sehr ausführlich darstellt und formelle Beschreibungen der Entscheidungskalküle beider Seiten liefert. Damit ist prinzipiell alles vorhanden, was zu einer Umsetzung ihres Ansatzes *'in software'* notwendig ist, was möglicherweise eine interessante Alternative zu den aus der Informatik kommenden Ansätzen darstellt. Aus Sicht der Soziologie ist an Rippergers Ansatz allerdings die mangelnde Betrachtung von Vertrauen als einem sozialen Phänomen zu kritisieren – Institutionen und institutionelle Rahmenbedingungen werden zwar als für die Bildung und Aufrechterhaltung von Vertrauen wichtiges Umfeld gesehen, Vertrauen bleibt aber auf das Entscheidungskalkül des *homo œconomicus* beschränkt. Dennoch – oder gerade deswegen – liegen hier möglicherweise Entwicklungspotenziale und Anregungen für die Einbettung eines ökonomisch orientierten Vertrauensmodells in ein Multiagenten-System.

4.1.3 Zusammenfassung agentenzentrierte, solitäre Ansätze

Fazit

Diese Kategorie wird vor allem durch Marshs Ansatz dominiert, der für viele weitere Entwicklungen prägend war. Neben diesem Ansatz wurde auf Bachmann und Rippberger eingegangen, die beide allerdings nicht direkt als implementierbare Modelle gedacht sind. Eine Hauptkritik an Marsh lag – neben der Frage, ob seine Formalisierungen der Einschätzung von Vertrauenswürdigkeit und Kooperationsschwelle tragfähig sind – in der fehlenden Einbeziehung des Urteils anderer Agenten in die Berechnung der Vertrauenswürdigkeit. Einige der unter anderem als Reaktion auf diese Schwachstelle in Marshs Modell entstandenen Ansätze bilden das Thema des nächsten Abschnitts.

4.2 Agentenzentrierte, soziale Ansätze

Abgrenzung

Wie die agentenzentrierten, solitären Ansätze erfolgt auch bei den agentenzentrierten, sozialen Ansätzen die Berechnung und Bewertung der Vertrauenswürdigkeit anderer überwiegend ‘im Agenten’ und nicht durch eine externe Instanz. Anders als bei den solitären Ansätzen, bei denen Agenten sich auf ihre eigenen Beobachtungen verlassen mussten, um andere Agenten zu bewerten, sehen die verschiedenen sozialen Ansätze (Abdul-Rahman / Halles 1997; Rasmusson 1996; Rasmusson / Janson 1996; Schillo 1999; Yu / Singh 2000) eine Einbeziehung der Aussagen Dritter in die Berechnung der Vertrauenswürdigkeit vor. Wie dies im einzelnen geschieht, welches Gewicht der Bewertung Dritter zugewiesen wird, und wie mit Bekannten von Bekannten umgegangen wird, unterscheidet die hier diskutierten Verfahren von einander. Allen Verfahren gemeinsam ist die Tatsache, dass die Erfahrungen anderer Agenten in den Prozess zur Berechnung der Vertrauenswürdigkeit bzw. zur Suche des besten Partners einbezogen werden können. Dabei spielt insbesondere die Frage eine Rolle, wie mit den Aussagen von Agenten umgegangen wird, die einem selbst als weniger vertrauensvoll erscheinen. Dies deutet schon darauf hin, dass alle sozialen Ansätze mit dem Problem konfrontiert werden, wie intersubjektive Reputationswerte dargestellt werden sollen.[20] Ein weiteres Problem liegt in der möglichst effizienten Speicherung des Wissens über andere Agenten, das ja jetzt nicht mehr nur die eigenen Erfahrungen mit anderen Agenten umfasst, sondern auch Bewertungen anderer Agenten, die sich möglicherweise mit der Zeit oder in Bezug auf unterschiedliche Situationen auch verändern.

4.2.1 Michael Schillo – *TrustNet* oder: Du bist nicht alleine

Schillo 1999

Michael Schillo (1999; Schillo et al. 1999) hat in Erweiterung eines Ansatzes von Castelfranchi et al. (1997) um eine Kommunikationskomponente ein Konzept für die

---

[20] Vgl. die Diskussion um die Nutzung eines quantitativen Maßes für Vertrauen in Bezug auf Marsh – nun geht es u.a. darum, wie ein und die selbe Maßzahl für Reputation oder Vertrauenswürdigkeit von zwei unterschiedlichen Agenten interpretiert wird.

Implementierung von Vertrauen in Multiagenten-Systeme entwickelt, das nicht nur eine fundierte Formalisierung von Vertrauen erlaubt – Schillo grenzt sich hier mit Castelfranchi et al. deutlich von Marsh ab (vgl. Schillo 1999: 25ff) –, sondern darüber hinaus auch die Möglichkeit bietet, dass Agenten untereinander Wissen über andere Agenten (Zeugenaussagen) austauschen können. Auch dabei wird die in offenen Systemen fehlende Gültigkeit der *benevolence assumption* berücksichtigt, indem Einschätzungen der Ehrlichkeit und des Grad an Altruismus bezogen auf andere Agenten in die Überlegungen miteinbezogen werden. Um sein Modell zu testen, hat Schillo beispielhaft ein um eine Auswahlphase erweitertes wiederholtes Gefangenendilemma gewählt. Prinzipiell ist das *TrustNet*-Modell aber für den Einsatz in offenen Systemen gedacht, etwa in Bezug auf den *eCommerce*-Bereich (vgl. Schillo 1999: 66ff; Schillo et al. 1999: 99f).

Szenario

Eine Runde des modifizierte Gefangendilemma-Spiel läuft wie folgt ab: (1) Bezahlen einer Teilnahmegebühr, (2) Verhandlungen, in denen die Spielpartner zusammenfinden – hier können Agenten (gelogene) Ankündigungen über ihre Intentionen machen, (3) Spiel einer Gefangenendilemma-Runde (beide beteiligten Agenten geben simultan ihre Spielzüge bekannt), (4) Bekanntgabe der Spielergebnisse – sichtbar nur für die Agenten in der Nachbarschaft eines beteiligten Agenten, (5) Auszahlung der Preise. In der Phase (2) haben Agenten die Möglichkeit, ihnen bekannte Agenten über Dritte zu interviewen, und auch die Spielergebnisse aus der Phase (4) können dazu genutzt werden, ihr *TrustNet* auf den neusten Stand zu bringen. Aufgrund dieser Mechanismen können Agenten, die ihre in Phase (2) gemachten Ankündigungen nicht eingehalten haben, relativ schnell nicht nur von ihren direkten Spielpartnern, sondern auch von deren Nachbarn ausfindig gemacht werden, was dazu führt, dass unehrliche Agenten nach einigen Runden keine Partner mehr finden und so auch keine Punkte verdienen können. (Schillo et al. 1999: 98ff).

Altruismus, Ehrlichkeit

Um die Vertrauenswürdigkeit anderer Agenten einschätzen zu können, bilden die einzelnen Agenten Modelle über deren Ehrlichkeit und deren Altruismus. Mit *Ehrlichkeit* bezogen auf einen Agenten Q ($E(Q)$) ist dabei gemeint, wie wahrscheinlich es ist, dass Interaktionen normgemäß, d.h. in diesem Fall so wie angekündigt ausgeführt werden (Anzahl der Interaktionen, die wie angekündigt ausgeführt werden, bezogen auf alle Interaktionen), während $A(Q)$ ein Maß für die Wahrscheinlichkeit ist, dass ein Agent sich *altruistisch* verhält (also ohne Rücksicht auf sein Gegenüber kooperiert). Mit diesen beiden Wahrscheinlichkeitsmaßen korrespondiert jeweils das Modell eines Agenten X über einen Agenten Q bezogen auf dessen Ehrlichkeit $E_X(Q)$ bzw. bezogen auf dessen Altruismus-Neigung $A_X(Q)$. (Schillo 1999: 49ff). Da es sich in all diesen Fällen um Wahrscheinlichkeiten handelt, bewegen sich diese Maße alle im Intervall [0;1].

TrustNet

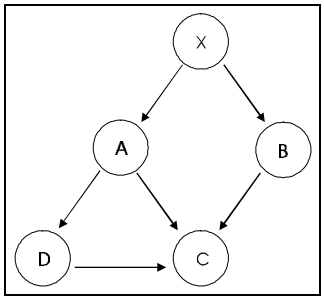


Abb. 4 – *TrustNet* nach Schillo (1999: 73)

Das *TrustNet* (Abb. 4) als zentrales Element von Schillos Ansatz dient als Speicher für die Einschätzung der anderen Agenten und hat die Form eines gerichteten Graphen. Die Knoten des Graphen repräsentieren Agenten, während die Kanten für Beobachtungen stehen. Beobachtungen sind Mengen von Tripeln der Form (*Spielrunde*, *alt*, *ehrl*), wobei *alt* (Altruismus) und *ehrl* (Ehrlichkeit) die Werte ja oder nein annehmen können oder aber auch verschwiegen werden können (Schillo 1999: 52). Jeder Agent hat seine eigene Datenstruktur, in der er durch den Wurzelknoten des Graphen (keine eingehende, nur ausgehende Kanten) repräsentiert wird, während die anderen Knoten für die – direkt (direkte Verbindung Wurzelknoten – Agentenknoten) oder über die Mitteilungen anderer (Zeugen) – beobachteten Agenten stehen. Die Mitteilungen werden dabei mit den Kanten gespeichert, die aus diesen abgeleiteten Einschätzungen werden als Modell der Ehrlichkeit und Modell des Altruismus an den jeweiligen Knoten abgelegt. (Schillo 1999: 72ff). Prinzipiell kann dieser Graph Zyklen enthalten (in Abb. 4 wäre dies der Fall, A→D→C um eine Mitteilung von C über A ergänzt würde); diese werden allerdings durch einen geeigneten Algorithmus (Schillo 1999: 79) so entfernt, dass möglichst die Kante mit dem größtmöglichen Informationsgehalt (bzw. könnte dies auch die aktuellste Kante sein) erhalten bleibt. Ein weiteres Problem stellt die Integration von Aussagen mehrerer Zeugen über einen Agenten dar. Schillo (1999: 77) stellt dazu einen rekursiven Algorithmus vor, der mit Hilfe von wahrscheinlichkeitstheoretischen Abschätzungen und Überlegungen zur Motivation der Agenten[21] eine integrierte Abschätzung der Ehrlichkeit bzw. des Altruismus eines Agenten liefert.[22]

Vertrauensberechnung

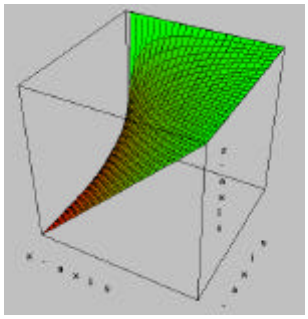


Abb. 5 – $V_X(Q)$: x→A(Q), y→E(Q), z→V(Q)

Ausgehend von den so erzeugten Modellen über die Ehrlichkeit und den Altruismus eines Agenten lässt sich jetzt dessen Vertrauenswürdigkeit V berechnen (unter der Voraussetzung, dass der Agent Q dem Agenten X eine Kooperation angegeben hat – Vertrauenswürdigkeit ist dann die Wahrscheinlichkeit dafür, dass diese Kooperation auch erfolgt). (Schillo 1999: 60)

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

Diese Formel lehnt sich an das Modell von Castelfranchi et al. (1997) an, wobei einige dort nur semantisch, aber nicht als Berechnungsmodi angegebene Werte von Schillo auf 1,0

---

[21] In diesem Szenario kann davon ausgegangen werden, dass Agenten andere Agenten in ein möglichst ungünstiges Licht stellen wollen und deswegen eher Aussagen über beobachteten Betrug machen als Aussagen über beobachtetes ehrliches Verhalten. Mit Hilfe dieser Überlegung lässt sich eine auf stochastische Methoden gestützte Heuristik zur Ermittlung der Zahl der von einem Agenten verschwiegenen Beobachtungen ehrlichen Verhaltens ableiten, die dann wiederum dazu genutzt werden kann, die Aussagen verschiedener Zeugen sinnvoll zu integrieren (vgl. für eine genauere Beschreibung Schillo 1999: 74-77).

[22] Eine einfachere Mittelwertsberechnung lässt sich wegen der teilweisen Überlappungen verschiedener Aussagen nicht.anwenden, so dass Schillo letztlich zu einem relativ aufwändigen Algorithmus kommt (vgl. zur Begründung, warum der Mittelwert nicht ausreicht, Schillo et al. 1999: 96ff bzw. Schillo 1999: 40ff).

gesetzt wurden. Schillo unterscheidet explizit zwischen der Vertrauenswürdigkeit *V*, und Vertrauen, womit die Haltung eines Agenten bezüglich einer Interaktion mit einem anderen Agenten gemeint ist (hier bezieht sich Schillo auf eine Vertrauens-Definition von Deutsch). Das Maß dieses Vertrauens entspricht dem aus subjektiver Sicht ermittelten Maß der Vertrauenswürdigkeit *V*. Abb. 5 zeigt, wie sich $V_X(Q)$ für unterschiedliche Werte von $A_X(Q)$ und $E_X(Q)$ verhält. Auffällig ist dabei, dass bei einer hohen Einschätzung für die Ehrlichkeit der Altruismus nur noch eine sehr geringe Rolle spielt.

Diskussion

Im Vergleich mit Marsh fällt auf, dass Schillo sehr viel mehr Wert auf die Herleitung und Begründung der von ihm eingesetzten Formeln legt. Da Schillo mit Wahrscheinlichkeiten für bestimmte Verhaltensweisen rechnet, gibt es keine Probleme mit der Multiplikation negativer Werte. Schillo geht von gesellschaftlichen Normen als Pendant zur Vertrauenswürdigkeit aus, was sich in der Verwendung von Altruismus und Ehrlichkeit (sowie in seiner Definition des Betrugs durch Verschweigen von Informationen) niederschlägt. Die Vertrauenswürdigkeit $V_X(Q)$ hat in etwa die Bedeutung, die das situative Vertrauen $T_X(Q, \alpha)$ bei Marsh einnimmt; sie wird im Spielszenario dazu genutzt, den vertrauenswürdigsten Kooperationspartner auszuwählen, indem – nach der Evaluation von Zeugenaussagen – der Agent mit der maximalen Vertrauenswürdigkeit ausgewählt wird. (Vgl. die bei Schillo 1999: 85ff angegebenen Algorithmen). Statt des einfacheren Gedächtnisses über situative Vertrauenswerte bei Marsh schlägt Schillo die Datenstruktur *TrustNet* vor, in der für bekannte oder über Zeugen bekannte Agenten Informationen über deren – angebliches – Verhalten in bestimmten Spielrunden sowie eine Einschätzung von Ehrlichkeit und Altruismus gespeichert wird. Auf die Einbindung des Situationsbezugs verzichtet Schillo; dieser würde *TrustNet* auch noch einmal deutlich verkomplizieren. Ein Äquivalent zu dem sich je nach Veränderung der Situation ändernden *basic trust* eines Agenten bei Marsh gibt es bei Schillo ebenfalls nicht.

Unbekannte Agenten

Auf die problematische, weil informationslose Situation eines neuen Agenten geht Schillo nicht ein. In der Phase (2) – Auswahl der Spielpartner – seines Szenarios werden jeweils alle anderen Agenten über die Selbstbeschreibung des Agenten informiert, der mit der Partnersuche dran ist. Diese können darauf mit Zustimmung oder Ablehnung reagieren; unter allen zustimmenden Agenten kann der suchende Agent dann denjenigen auswählen, der ihm am vertrauenswürdigsten erscheint oder andere Agenten befragen. Es ist allerdings unklar, was passiert, wenn ein Agent noch zu keinem anderen Agenten ein Modell gebildet hat – wie werden die Erwartungen an Ehrlichkeit und Altruismus initialisiert? Auch das soziale Element in Schillos Modell hilft hier nicht unbedingt weiter, da Zeugenaussagen ja wiederum mit der Einschätzung der Ehrlichkeit der Zeugen verrechnet werden – die aber noch nicht definiert ist. Allerdings dürfte sich dieses Anfangsproblem recht schnell lösen, sobald nur einige wenige Erfahrungswerte vorliegen, da *TrustNet* Agenten dazu in die Lage versetzt, Zeugenaussagen von Zeugen, über deren Glaubwürdigkeit sie wiederum nur von

Dritten gehört haben, ohne Probleme in das Agentenmodell einzubauen. Da die Kommunikation zwischen den Agenten sich nur auf Beobachtungen (*X* war ehrlich / unehrlich) und nicht auf Einschätzungen bezieht, umgeht Schillo die Frage der globalen Kompatibilität von Vertrauenseinschätzungen. Theoretisch könnten *TrustNet*-Agenten mit ihnen völlig unbekannten Agenten kooperieren, wenn nur der Austausch von Beobachtungen über Altruismus und Ehrlichkeit über Dritte gegeben ist.

Nutzbar für eCommerce? Sowohl die theoretischen Überlegungen als auch die konkrete Implementierung von *TrustNet* bezieht sich bei Schillo auf das *offen gespielte Gefangenendilemma mit Partnerauswahl*. Es stellt sich also die Frage, wieweit der Anspruch, dass das Modell auf andere Bereiche übertragbar ist, eingelöst werden kann. In einem Vergleich zwischen offen gespieltem Gefangenendilemma mit der Situation auf virtuellen Märkten sieht Schillo (1999: 66) einen direkten Zusammenhang zwischen beiden Situationen. Allerdings gibt es – unabhängig von der Frage, wieweit die theoretischen Annahmen nicht doch nur für das konkrete Szenario gelten – keine Aussagen dazu, wie sich *TrustNet* bei völlig ungleichen Informationsbedingungen verhält (einige Agenten sind schon lange am Markt, andere kommen neu hinzu), und ob die Skalierbarkeit für große Märkte gegeben ist (potenziell ja mehrere 1000 bis 100.000 Agenten). Das Argument der mangelnden Skalierbarkeit verliert etwas an Bedeutung, wenn berücksichtigt wird, dass sich auch auf großen Märkten möglicherweise nur kleine Cliquen einander bekannter Agenten herausbilden, was allerdings zur Frage führt, wie auf einem offenen Marktplatz das Problem der Nachbarschaft bzw. Bekanntheit gelöst werden soll. Prinzipiell scheint *TrustNet* – dessen Vorteile gegenüber Agenten ohne 'Vertrauensschutz' Schillo im empirischen Teil seiner Studie zeigt (1999: 92ff) – allerdings bei der Wahl geeigneter Kriterien und Kommunikationsregeln auf andere Szenarien, insbesondere auch auf agentenbasierte Märkte, übertragbar zu sein.

Sicheres Gesamtsystem? Nach Aussage von Schillo »sinkt in sozial kompetenteren Gesellschaften unter Verwendung von Vertrauen die Performanz für die Egoisten.« (1999: 102). Oder anders gesagt: längerfristig – sobald sich ausreichend genaue Modelle über die anderen Agenten herausgebildet haben – schneiden altruistischere (und ehrlichere) Agenten besser als ihre weniger freundlichen Gegenparts ab. Auf den ersten Blick scheint dieser Ansatz also tatsächlich starke Anreize dafür zu schaffen, ehrlich und kooperativ (d.h. altruistisch in diesem Spiel) miteinander umzugehen. Unklar ist allerdings, ob diese Eigenschaften auch auf einem realen Markt wiederzufinden sind, in dem ja immer wieder neue Agenten dazu kommen und alte abwandern, oder ob zwar die Eigenschaft starker Anreize für ehrliches Verhalten gegeben ist, dies aber letztlich zu einer Bildung 'kooperativer Cliquen' führt, so dass neue Agenten keine Chance haben, am Handel beteiligt zu sein – außer am letztlich kontraproduktiven Handel mit egoistischen, nicht an den 'kooperativen Cliquen' beteiligten Agenten. Damit stellt sich die Frage, ob eigenes und soziales Wissen über andere Agenten ausreicht, oder ob nicht noch zusätzliche Möglichkeiten – auch im Sinne des von Sztompka

(1999) genannten Hinweisbündels für Reputation – in die Berechnung der Vertrauenswürdigkeit eingeführt werden sollten.

Gültigkeit der Annahmen

Schillo selbst geht kurz auf soziologische (sozionische) Aspekte des von ihm gewählten Ansatzes ein (1999: 68ff). Seine Interpretation von Vertrauen als Handlungskoordinationsmechanismus zur Reduktion von Risiko koppelt er dabei einerseits an Luhmann (vgl. Luhmann 1989) als auch an die sozialpsychologischen Vertrauensdefinitionen von Morton Deutsch (1973). Mit Bachmann (vgl. Bachmann 1999) beschreibt Schillo die konkrete Funktion dieses Vertrauensmechanismus als die Möglichkeit, die Komplexität des Handlungssystems über »spezifische Annahmen über das zukünftige Verhalten des Gegenübers« (Schillo 1999: 69) zu reduzieren. Diese Funktion sieht er durch die Einbeziehung von Zeugenaussagen in den Modellbildungsprozess gegeben. Er stellt weiterhin die ausgrenzenden Aspekte in Bezug auf betrügerische Agenten dar; auch dies kann durchaus in Relation zu den Annahmen der verschiedenen soziologischen Theorien gesetzt werden, etwa zu Luhmanns Gesetz des Wiedersehens und der daraus abgeleiteten Notwendigkeit für Agenten, sich vertrauenswürdig zu zeigen und nicht zu betrügen. Im Vergleich mit Luhmanns Theorie über Vertrauen erweist sich *TrustNet* allerdings als deutlich unterkomplex, was sich etwa in der Behandlung des Themas Macht zeigt. Im Vergleich zu Marsh stellt die fehlende Kontextabhängigkeit der Agentenbewertungen möglicherweise ein Problem dar – auch in den von Schillo zugrundegelegten Theorien wird davon ausgegangen, dass Vertrauen nicht einfach von einem Kontext auf alle Aspekte einer Person übertragen werden kann, sondern dass es hierfür Grenzen gibt. Schließlich bleibt wie bei Marsh die – allerdings durch die realitätsnähere Formalisierung abgemilderte – Kritik am rein quantitativen Vorgehen.

Technische Kriterien

Schillo gibt die Speicherkomplexität je Agent mit $O(n^2\ r)$ an, wobei n für die Zahl der diesem Agenten bekannten anderen Agenten und r für die Zahl der gespielten (bzw. gespeicherten) Runden steht. An dieser Stelle fällt auf, dass eine bei Marsh auch theoretisch begründete Begrenzung des Speichers wegfällt; diese lässt sich aber leicht nachträglich in *TrustNet* einbauen (etwa so, dass am wenigsten genutzte Kanten nach einer gewissen Zeit verblasen[23]). Die Speicherkomplexität ist deutlich größer als bei Marsh (in der hier verwendeten Notation: $O(nr)$), was sich damit begründen lässt, dass zusätzlich zu den Modellen der anderen Agenten auch die Kanteninformationen gespeichert werden müssen. Was die Zeitkomplexität angeht, verwendet *TrustNet* das Prinzip der *lazy evaluation*, d.h. Neuberechnungen der Einschätzungen werden erst vorgenommen, wenn konkrete Anfragen vorliegen. Das Einfügen von Kanten in *TrustNet* erfolgt in kubischer Zeit ($O(n^3)$), ist also

[23] Eine derartige Erweiterung von *TrustNet* hätte auch den Vorteil, dass Agenten in ihrer Modellbildung flexibler auf Veränderungen im Verhalten anderer Agenten eingehen könnten, da ja, anders als in Schillos Test-Szenario, in der Realität eines offenen Marktes nicht davon auszugehen ist, dass die Wahrscheinlichkeit ehrlichen / altruistischen Verhaltens über die Zeit gleich bleibt.

relativ aufwändig und ein die Skalierbarkeit des Ansatzes begrenzender Faktor. Das Einfügen eines Knotens oder das Auslesens eines bereits berechneten Wertes erfolgt in $O(n \log n)$. Eine komplette Neuberechnung des Netzes – wenn alle Kanten des Wurzelknoten sich verändert haben – hat die Zeitkomplexität $O(n^2 r)$. (Schillo 1999: 79f).

Fazit

TrustNet stellt einen interessanten Ansatz zur Vertrauensberechnung dar, der etwas aufwändiger als Marshs Ansatz ist, dafür aber sowohl bei der Art der Berechnung der Vertrauenswürdigkeit als auch aufgrund der Einbeziehung von Interagenten-Kommunikation fundierter erscheint. Ob die gewünschte Funktionalität – die Steigerung von Kooperationsbereitschaft und die Reduktion betrügerischen Verhaltens – auch abseits des doch relativ geordneten, synchronisierten Testszenarios erreicht wird, bleibt allerdings fraglich. Für eine Übertragung auf große offene Systeme wie elektronische Marktplätze lässt *TrustNet* noch einige Fragen sowohl technischer Art als auch in Bezug auf das Verhältnis insbesondere zwischen *greenhorns* und erfahrenen Agenten offen.

4.2.2 Weitere agentenzentrierte, soziale Ansätze

Rasmuson 1996

Lars Rasmusson (1996) und Sverker Janson (Rasmusson / Janson 1996; Rasmusson et al. 1997) stellen einige allgemeine Überlegungen zur Sicherheit agentenbasierter, offener Marktplätze an, die im Rahmen eines einfachen Simulationssystems überprüft werden. Sie plädieren – vor allem auch aus Gründen der Skalierbarkeit, aber auch, um eine Monopolsituation zu vermeiden – dabei für einen dezentral umgesetzten, reaktive Sicherheitsansatz. Sicherheit soll also nicht oder nur ergänzend durch zentrale Institutionen vermittelt werden, primär aber durch das Verhalten der einzelnen Agenten entstehen. Dafür skizzieren sie – neben Vorschlägen, etwa *Trusted Third Parties* für Finanztransaktionen heranzuziehen, um so das dabei entstehende Risiko zu minimieren, oder spezielle *Reviewer Agents* einzuführen – insbesondere auch einen sozialen Ansatz: Dabei sollen Agenten in der Lage dazu sein, aktiv Informationen über ihre Erfahrungen mit anderen Agenten bzw. über diese gehörte Gerüchten weiterzugeben (*Gossip*, also 'Tratschen'). Dabei entsteht allerdings das auch von Schillo beschriebene Problem, dass Agenten – zumindest in einem stark kompetitiv ausgerichtetem Rahmen – einen gewissen Anreiz haben, zu lügen, da Informationen über das Verhalten anderer Agenten Wettbewerbsvorteile für die Konkurrenz bietet (vgl. Rasmusson / Janson 1996, 14). Eine mögliche Erweiterung des 'Tratschen' könnte dazu führen, das Agenten sich selbst anpreisen (*Advertising*, vgl. Rasmusson 1996: 13f). Ein möglicher Umgang mit dem Problem der absichtlichen Weitergabe falscher Informationen könnte in einer Mediatisierung durch Geld liegen, d.h. dass ein Agent *A*, um sich selbst oder einen anderen Agenten (*B*) anzupreisen, dafür zahlt, dass Agent *C* sich an ihn bzw. an *B* erinnert (vgl. Rasmuson 1996: 13, 25).

Interessant sind auch die Ergebnisse der Marktsimulation (Rasmusson 1996: 29ff), obwohl hier nur solitäre Ansätze (*favourite choice*) in die Untersuchung einbezogen wurden. Dabei

zeigte sich, dass eine reine Ausrichtung am Preis die Gesamtqualität des Marktes reduziert, und stabile Zustände erst nach dem Bankrott vieler Kauf-Agenten erreicht werden. Wird die Ausrichtung am Preis mit der Beschränkung auf einen Marktplatz kombiniert, entstehen Monopolsituationen. Das Hinzuziehen der *favourite-choice*-Strategie verringert den Anteil betrügerischer Agenten und führt kombiniert mit dem Mehrere-Marktplätze-Ansatz zu einem stabilen, qualitativ hochwertigen Zustand. Rasmusson vermutet, dass eine Einbeziehung eines auf der Weitergabe von Gerüchten basierenden Reputationsmechanismus weitere Verbesserungen mit sich bringen würde.

Abdul-Rahman/Halles 1997

Einen etwas anderen Ansatz als Schillo verfolgen Alfarez Abdul-Rahman und Stephan Halles (1997). Ihr direkt auf den internetbasierte Multiagenten-Systeme bezogenes Modell soll dazu diesen, Vertrauen als Grundlage von informellen oder kurzfristigen Beziehungen oder in Bezug auf kommerzielle Ad-hoc-Transaktionen zu implementieren. Dazu schlagen sie vor, dass jeder Agent ein Netzwerk von *trust relationsships* in seiner Datenbank mitführt. Eine *trust relationship* ist dabei definiert als eine Beziehung zwischen genau zwei Entitäten, die gerichtet ist, und unter Umständen transitiv sein kann. Dabei unterscheiden sie zwischen *direct trust relationships* (»Alice vertraut Bob«) und *recommender trust relationships* (»Alice vertraut den Empfehlungen, die Bob zur Vertrauenswürdigkeit anderer ausspricht«). Ein interessanter Unterschied zu den anderen vorgestellten Formalisierung liegt darin, dass Abdul-Rahman / Halles (1997: 53) aufgrund der qualitativen Natur von Vertrauen nicht mit Wahrscheinlichkeitswerten oder einem Intervall [-1;1) arbeiten, sondern mit diskreten Werten (vgl. Tabelle 8), die jeweils auch nicht allgemeingültig sind, sondern auf bestimmte *trust categories* (»Alive vertraut Bob, wenn es um den Handel mit Tischen geht.«) bezogen sind. Reputation definieren Abdul-Rahman / Halles (1997: 54) dementsprechend als Tripel (*ID-Agent*, *Trust-Category*, *Trust-Value*) definiert. Jeder Agent speichert derartige Reputationen in seiner eigenen Datenbank und nutzt diese, um Empfehlungen an andere auszusprechen. Darüber, wie ein Agent zu diesen Bewertungen kommt, machen Abdul-Rahman / Halles allerdings keine Aussage. Den Kern ihres Papers stellen Überlegungen zu einem *Recommendation Protocol* dar, mit dessen Hilfe Anfragen für Empfehlungen, Empfehlungen selbst und Aktualisierungsanfragen innerhalb eines Multiagenten-Systems kommuniziert werden können. Eine Empfehlungsanfrage wird solange weitergereicht, bis ein Agent (oder mehrere Agenten) gefunden wird, der oder die eine Empfehlung zur gefragten Kategorie geben kann, und dem / denen der vorletzte Agent in der Kette vertraut. Ausgehend davon kann der anfragende Agent dann mit folgender Formel[24]

[24] *tv*(*Rx*) ist dabei der *recommender trust value* der verschiedenen an der Empfehlung beteiligten Agenten und *rtv*(*T*) ist der vom letzten Agenten ausgesprochene empfohlene Vertrauenswert. Wenn ein Agent mehr als eine Empfehlung über einen anderen Agenten erhält, werden diese von ihm gemittelt.

Wert	Bedeutung für *direct trust relationship*	Bedeutung für *recommender trust rel.*
-1	*Distrust* – Nicht vertrauenswürdig	*Distrust* – Nicht vertrauenswürdig
0	*Ignorance* – Keine Bewertung möglich	*Ignorance* – Keine Bewertung möglich
1	*Minimal* – Niedrigstmögliches Vertrauen	'Nähe' der Beurteilung des Empfehlenden zur eigenen Beurteilung von Vertrauenswürdigkeiten
2	*Average* – Durchschnittliches Vertrauen– den meisten Entitäten entspricht dieser Level	
3	*Good* – Vertrauenswürdiger als *Average*	
4	*Complete* – Völliges Vertrauen zu dieser Entität	

Tabelle 8 – Diskrete Vertrauenswerte nach Abdul-Rahman / Halles (1997: 53)

$$tv_r(T) = \frac{tv(R_1)}{4} \times \frac{tv(R_2)}{4} \times \ldots \times \frac{tv(R_n)}{4} \times rtv(T)$$

die Vertrauenswürdigkeit einer Empfehlung berechnen. Sowohl die – von den Autoren selbst zugegebene – fehlende Begründung für die spezielle Art der Reputationsberechnung als auch das Fehlen von Überlegungen dazu, wie Agenten zu ihren direkten *trust values* kommen, als auch die Notwendigkeit der global standardisierten Kategorien-Ontologie stellen ernsthafte Probleme dieses Ansatzes dar. Dennoch scheint er mir als Skizze eines qualitativen und zugleich algorithmisch ausgeführter Ansatzes auf jeden Fall erwähnenswert. Dazu gehört insbesondere auch die Überlegung, Misstrauen (bezüglich einer bestimmten Kategorie) nicht als ein Kontinuum, sondern als einen Zustand zu konzipieren.

Yu / Singh 2000

Ein konkretes, in einen Punkten – etwa bei der Realisierung von Vertrauen durch das Intervall [-1;1] – stärker an Marshs Formalismus angelehntes Modell des agentenzentrierten, sozialen Reputationsmanagements[25] für elektronische Gemeinschaften im allgemeinen und Multiagenten-Systeme im besonderen legen Bin Yu und Munindar P. Singh (2000) vor. Zum einen sehen sie dabei wie Rasmusson ebenfalls einen *Gossip*-Mechanismus vor (»If an agent *A* encounters a bad partner *B* during some exchange, A will penalize B by decreasing ist rating of B by β and informing its neighbors.« (Yu / Singh 2000: 6)). Tratsch soll dabei inkrementell, von Nachbarschaft zu Nachbarschaft durch das Netzwerk der Agenten weitergegeben werden. Zum anderen sehen sie vor, dass es einen Mechanismus geben soll, mit dem die Aussagen anderer Agenten (*Zeugen*) in die eigene Berechnung der Vertrauenswürdigkeit einbezogen werden. Ähnlich wie Abdul-Rahma / Halles sollen dazu *referral chains* gebildet werden, über die, möglicherweise über viele Zwischenstationen – hier allerdings quantitativ angelegte – Zeugenaussagen dem anfragenden Agenten zur Verfügung gestellt werden. Für die Veränderung des Vertrauens nach guten oder schlech-

[25] Reputation wird hier mit Vertrauen(swürdigkeit) gleichgesetzt.

Veränderung des Vertrauens nach einer Interaktion[26]

$T_i(j)^t > 0$, j hat kooperiert	$T_i(j)^{t+1} = T_i(j)^t + \alpha\,(1-T_i(j)^t)$
$T_i(j)^t < 0$, j hat kooperiert	$T_i(j)^{t+1} = (T_i(j)^t + \alpha) / (1-\min(\lvert T_i(j)^t\rvert, \lvert\alpha\rvert))$
$T_i(j)^t = 0$, j hat kooperiert	$T_i(j)^{t+1} = \alpha$
$T_i(j)^t > 0$, j hat betrogen	$T_i(j)^{t+1} = (T_i(j)^t-\beta) / (1-\min(\lvert T_i(j)^t\rvert, \lvert\beta\rvert))$
$T_i(j)^t < 0$, j hat betrogen	$T_i(j)^{t+1} = T_i(j)^t + \beta\,(1-T_i(j)^t)$
$T_i(j)^t = 0$, j hat betrogen	$T_i(j)^{t+1} = \beta$

Einbeziehung von Zeugenaussagen über Agent n

L = Anzahl unterschiedlicher Zeugenaussagen in $E = \{E_{1w}, \ldots, E_{Lw}\}$[27]
V : die Teilmenge von E, die nur die Zeugenaussagen vertrauenswürdiger Zeugen, und von den gleichen Zeugen nur die beste Zeugenaussage enthält
$\hat{E}$ = Mittelwert aller Zeugenaussagen in V

$T_i(n)^t > 0$ und $\hat{E}>0$	$T_i(n)^{t+1} = T_i(n)^t + \hat{E}\,(1 - T_i(n)^t)$
$T_i(n)^t < 0$ xor $\hat{E}<0$	$T_i(n)^{t+1} = T_i(n)^t + \hat{E} / (1 - \min(\lvert T_i(n)^t\rvert, \lvert\hat{E}\rvert))$
$T_i(n)^t < 0$ und $\hat{E}<0$	$T_i(n)^{t+1} = T_i(n)^t + \hat{E}\,(1 + T_i(n)^t)$

Einbeziehung von Gerüchten ($T_k(n)$), die Agent i von k über n gehört hat

$T_i(n)^t > 0$ und $T_i(k)^t > 0$	$T_i(n)^{t+1} = T_i(n)^t + T_i(k)^t\,T_k(n)\,(1 - T_i(n)^t)$
$T_i(n)^t < 0$ und $T_i(k)^t < 0$	$T'i(n)^{t+1} = T_i(n)^t + T_i(k)^t\,T_k(n)\,(1 + T_i(n)^t)$
unterschiedliche Vorzeichen	$T'i(n)^{t+1} = (\,T_i(n)^t + T_i(k)^t\,T_k(n)\,) / (1-\min(\lvert T_i(n)^t\rvert, \lvert T_i(k)^t\,T_k(n)\rvert)$

Tabelle 9 – Berechnung der Vertrauenswürdigkeit nach Singh / Yu (2000)

ten Interaktionen, für die Einbeziehung von Zeugenaussagen und Tratsch geben Yu / Singh (2000) jeweils mathematische Formalisierungen an, die auch die besonderen Eigenschaften negativer Vertrauenswerte berücksichtigen (vgl. Tabelle 9). Im Rahmen umfangreicher Experimente zeigen Yu / Singh (2000: 7ff), dass ihr Modell bei der geeigneten Wahl von α (Zunahme des Vertrauens bei Kooperation) und β (Abnahme bei Betrug) zu einem stabilen System führt, bei dem die Reputation betrügerischer Agenten schnell abnimmt, während neue, kooperative Agenten eine zwar langsam, aber dafür nahezu linear wachsende Reputation erhalten. Die Frage, wie eine agentenzentrierte Speicherung des sozialen Wissens vorgenommen werden soll – etwa der *referral chains* – diskutieren Yu / Singh (2000) nicht.

---

[26] $T_i(j)^t$ – Vertrauen von i in j zum Zeitpunkt t,. $\alpha > 0$ – Zunahmefaktor, $\beta < 0$ – Abnahmefaktor von Vertrauen. [$\alpha| < |\beta|$, (z.B. α=0.05, β=-0.3), analog zur Annahme, dass Vertrauen schwer hergestellt und leicht zerstört werden kann.

[27] Berechnungen der Zeugenaussagen erfolgen über die jeweiligen Vertrauensketten, wobei Verzweigungen in der Kette jeweils die Kette, die der Agent an der Verzweigung für vertrauenswürdiger hält, weiterverfolgt wird, und die einzelnen Vertrauenswerte $T_i(j)$ (das Vertrauen des Zeugen in das Ziel, das Vertrauen des vorletzten in der Kette in den Zeugen usw.) mit Hilfe eines Operators $\otimes$, der wie folgt definiert ist:

$$x \otimes y: \begin{cases} xy, \text{ für } x \geq 0 \wedge y \geq 0 \\ -|xy|, \text{ sonst} \end{cases}$$

4.2.3 Zusammenfassung agentenzentrierte, soziale Ansätze

Fazit

Das Problem solitärer Ansätze, dass dort das Wissen eines Agenten über betrügerischeres Verhalten zwar diesem nützt, aber anderenAgenten und damit auch dem Marktplatz insgesamt nicht weiterhilft, wird in agentenzentrierten, sozialen Ansätzen dadurch gelöst, dass Agenten in die Berechnung ihrer Vertrauenswürdigkeit Informationen anderer Agenten einbeziehen. Dabei stellt sich die Frage, wie die Vertrauenswürdigkeit dieser Zeugenaussagen zu beurteilen ist. Schillo (1999) stellt ein Modell vor, dass aufgrund von stochastischen Methoden die Annahmen von Zeugen über die Wahrscheinlichkeit ehrlichen und altruistischen Verhaltens in die eigene Einschätzung einbezieht. Yu / Singh (2000) ergänzen ein eher an Marsh orientiertes Modlel von Vertrauenswürdigkeit eine soziale Komponente. Neben diesen eher quantitativen Formalisierungen schlagen Abdul-Rahman / Halles (1997) ein qualitativeres Modell für die Vertrauenswürdigkeit der Empfehlungen anderer vor. Die Einbeziehung des Wissens anderer Agenten in die eigenen Berechnungen trägt insgesamt dazu bei, das Ziel, Interaktionen mit betrügerischen Agenten einzustellen, schneller zu erreichen. Die Annahme, dass Agenten Wissen über andere Agenten weitergeben, entspricht realweltlichen sozialen Hinweissystemen für Reputation (vgl. Kapitel 3), ist allerdings unter ökonomischen Gesichtspunkten eher unwahrscheinlich. Rasmusson (1996) diskutiert Lösungen für dieses Dilemma.

4.3 ‘Objektive’ externe Bewertungsagenturen

Abgrenzung

Neben den agentenzentrierten Ansätzen existieren verschiedene Versuche, Reputation durch Dritte bewerten zu lassen. Dabei können diese Dritte sowohl andere, um NutzerInnen ihres Bewertungsdienstes konkurrierende Agenten (vgl. etwa den Vorschlag, *Reviewer Agents* einzuführen, bei Rasmusson 1996) als auch von Agenten nutzbare, zentralisiert vom System bereitgestellte Dienste sein. Analog zu der Unterscheidung zwischen solitären und sozialen agentenzentrierten Ansätzen möchte ich zwischen ‘objektiven’ und ‘subjektiven’ externen Bewertungsagenturen unterscheiden. ‘Objektiv’ soll sich dabei darauf beziehen, dass die Bewertungsagentur selbst – etwa anhand eines Standards für Qualitätskriterien – Bewertungen durchführt, ‘subjektiv’ meint, dass die Bewertungsagentur Urteile einzelner Agenten über andere Agenten sammelt und zusammenfasst. Da es nur relativ wenige ‘objektiv’ externe Ansätze gibt, soll diese Form des Vertrauensmanagements hier nur kurz anhand zweier Beispiele angerissen werden.

Rasmusson 1996

In Abschnitt 4.2.2 wurde schon dargestellt, welche Ideen Lars Rasmusson (1996) für die soziale Kontrolle virtueller Märkte entwickelt hat. Eine von ihm angedachte Komponente für das Reputationsmanagement derartiger Märkte sind sogenannte *Reviewer Agents* (vgl. Rasmusson 1996: 16ff). Diese sollen eine vergleichbare Funktion etwa zu Restaurantkritikern in der realen Welt übernehmen und gezielt andere Agenten – etwa durch anonyme Probekäufe – zu testen, und das entstandene Wissen über diese anderen Agenten zur

Verfügung stellen. Anders gesagt: *Reviewer* handeln mit Informationen über andere Agenten. Wichtig ist dabei natürlich zum einen die Reputation der *Reviewer*, zum anderen, wie erwähnt, das Verheimlichen ihrer Identität gegenüber getesteten Agenten, da diese sonst möglicherweise beeinflusst sein könnten, in diesem Fall doch ehrlich zu sein, doch nicht zu betrügen, oder doch einen niedrigeren Preis zu verlangen und eine höhere Qualität zu liefern. Möglicherweise müsste ein *Reviewer* in einer realen Implementation, in der Anonymität nicht gewünscht ist, aus einem System mehrerer miteinander kooperierender Agenten bestehen, wobei ein Agent für die Weitergabe von Informationen an andere da ist, während andere, vielleicht nur temporär existierende Agenten die Testkäufe erledigen. Ob *Reviewer* in einem agentenbasierten Marktsystem überleben können, in dem der Aufenthalt auf dem Handelsplatz Gebühren kostet, müsste überprüft werden – gerade in der beschriebenen Konstellation mehrerer verbundener Agenten könnte es möglich sein, dass der Verkauf von Reputationsdaten für das finanzielle Überleben des Agenten nicht ausreicht. Auch ist zu überlegen, für welche *Produkte* und welche Märkte *Reviewer* geeignet sind. Es liegt nahe, *Reviewer Agents* nicht nur eigene Tests durchführen zu lassen, sondern auch die Erfahrungen anderer auszuwerten (vgl. Kapitel 4.4).

Kuhlen 1999

Nicht aus dem Bereich der Agentensysteme kommt die Idee, das Vertrauensmanagement an (u.U. konkurrierende) darauf spezialisierte Institutionen zu übergeben. Rainer Kuhlen (1999a; 1999b: 324ff) beschreibt dies unter dem Titel des »Delegierten Vertrauensmanagements« und nennt als Beispiel die u.a. mit der *Electronic Frontier Foundation*, aber auch mit *CommerceNet* kooperierende Organisation *TRUSTe*[28]. Diese hat sich dass Ziel gesetzt, elektronischen Handel unter *Privacy*-Gesichtspunkten vertrauenswürdiger zu machen und vergibt ein auf den jeweiligen Websites anzubringendes Siegel an Firmen, die offen legen, wie sie mit den Daten von Kunden und Kundinnen umgehen (Weitergabe an Dritte etc.). Ich führe *TRUSTe* deswegen hier auf, weil hier das Prinzip einer Überprüfung durch vertrauenswürdige Dritte nach ‘objektiven’ Qualitätsstandards im Bereich des elektronischen Handels implementiert wird. Eine ähnliche Methode der Vertrauenssicherung lässt sich auch in Multiagenten-Systemen etablieren. Durch eine *TRUSTe* ähnliche Organisation geprüfte Agenten – etwa auf einen Betrug nicht ermöglichenden Quelltext, auf die Authentizität ihrer angeblichen Eigentümer, auf die Einhaltung von Sicherheitsstandards bei der Datenübertragung könnten sich dann mit einem Siegel schmücken, dass auf Anfrage potenziellen Partnern gezeigt werden würde, was diese in ihre Einschätzung der Reputation einbeziehen könnten. Alternativ könnte auch ein die Qualitätsstandards vergebende Organisation repräsentierender Agent Anfragen über Dritte entgegennehmen und mit Vertrauenseinschätzungen beruhend auf der externen Prüfung antworten.

---

[28] Siehe dazu auch die Selbstdarstellung von *TRUSTe*, http://www.truste.org/about/about_whitepaper.html.

4.4 'Subjektive' externe Bewertungsagenturen

Abgrenzung

Im Gegensatz zu den 'objektiven' externen Bewertungsagenturen setzen 'subjektive' Agenturen bei ihrer Bewertung auf die gesammelten Erfahrungen anderer, die zu einer Bewertung zusammengefasst werden. Natürlich lässt sich dies mit anderen Ansätzen kombinieren. Wenn es um menschliche NutzerInnen geht, dienen derartige Ansätze häufig dazu, schon vorliegende soziale Netze wiederzugeben.

eBay

Eines der Vorbilder (vgl. Zacharia et al. 1999: 2) für eine zentralisierte, aber trotzdem 'subjektive' Reputationsdatenbank im Bereich des *eCommerce* ist das Online-Auktionshaus *eBay* [http://www.ebay.de]. Zu allen an Transaktionen beteiligten TeilnehmerInnen von *eBay* kann deren Reputation abgefragt werden. Diese besteht aus einem Reputationskoeffizienten – anfänglich 0, aus der Zahl der positiven und negativen Bewertungen sowie aus möglichen Text-Kommentaren. Auch etwa in Auktionsverläufen sind die einzelnen BieterInnen mit Sternen als Symbol für deren Reputation gekennzeichnet. Die Reputation bei *eBay* entsteht dadurch, dass bei Abschluss einer Transaktion beide TeilnehmerInnen eindringlich[29] gebeten werden, den anderen Transaktionspartner zu bewerten. Dies geschieht zum einen in Form des Feedbacks -1, 0 oder +1, und zum anderen in Form eines kurzen Textes Ähnliche Reputationsmanagementsysteme sind inzwischen auch bei vielen anderen Anbietern von Online-Geschäften mit pseudonymen KundInnen im Einsatz und scheinen deutlich zur Vertrauensbildung beizutragen – bis hin zur Politiksimulation *dol2day* [http://www.dol2day.de], die ebenfalls im Profil jeder TeilnehmerIn anzeigt, wie viele andere dieser Person vertrauen und misstrauen. Als Anreiz für die Teilnahme am Vertrauenssystem wird hier übrigens eine Kopplung der in der Politiksimulation erarbeiteten Punktezahl u.a. an die einem Vertrauenden gekoppelt. Um den Missbrauch der Vertrauensfunktion zu reduzieren, kostet das Aussprechen von Vertrauen / Misstrauen Punkte.

Zacharia 1999

Diese Idee – eine zentrale Instanz verwaltet die gesammelten Erfahrungen der einzelnen NutzerInnen in Form von Reputationswerten – lässt sich natürlich auch auf den Bereich der Multiagenten-Systeme übertragen. Das bekannteste Beispiel dafür ist Giorgos Zacharias System *Sporas / Histos* (1999; Moukas et al. 1999; Zacharia et al. 1999), das unter anderem für die Implementation eines *Better Business Bureau*-Service im MIT-System *Kasbah* (vgl. Maes et al. 1998) genutzt wurde. *Sporas* stellt dabei einen *eBay*-ähnlichen Mechanismus zur Verfügung, der aufgrund der paarweisen Bewertung nach Ablauf von Transaktionen globale Reputationswerte bereitstellt, die Teil der Identität der Agenten sind und von diesen nicht geändert werden können. *Histos* hingegen baut auf sozialen Netzwerken auf, um einen personalisierten Reputationswert zu berechnen, und lehnt sich damit an

---

[29] »Nehmen Sie Ihr Urteil und das Urteil der anderen bitte sehr ernst, denn Vertrauen ist das grösste Gut der eBay Gemeinschaft.« (Zitat aus dem Ratingsystem von http://www.ebay.de).

Web of Trust (vgl. dazu Kuhlen. 1999b 368ff) an. Beides geschieht nicht agentenspezifisch, sondern systemweit; d.h. wenn ein Agent das System nach der Reputation eines anderen Nutzers fragt, berechnet dieses – je nach dem Grad, mit dem dieser andere Nutzer in soziale Netze einbezogen ist – dessen Reputationswert und gibt dieses an den Agenten zurück. Ein Vorteil davon liegt darin, dass immer sämtliche Informationen bekannt sind, und einzelne Agenten sich nicht mit lügenden Agenten herumschlagen müssen, ein Nachteil dieser Zentralisierung des Reputationsservices liegt darin, dass ein zentrales System möglicherweise sehr viele Reputationsberechnungen durchführen muss, und das andere Algorithmen zur Reputationsberechnung nicht unterstützt werden. Wie aber sieht *Sporas / Histos* nun genau aus?

Sporas

Sporas dient dazu, für die NutzerInnen einer agentenbasierten, lose verbundenen *Online-Community* Reputationsbewertungen zur Verfügung zu stellen (vgl. Zacharia et al. 1999: 3f). Grundsätze dabei sind, dass neue Agenten[30] mit der Bewertung 0 anfangen, das kein Agent je unter diese Bewertung sinken kann (da sonst ein zu großer Anreiz gegeben wäre, einfach eine neue Identität anzulegen, wenn die Bewertung einmal unter 0 sinkt), dass jeweils nur die letzte paarweise Bewertung zwischen zwei Agenten berücksichtigt wird, und dass die Bewertung sich bei Agenten mit hoher Reputation weniger stark verändert als bei Agenten mit niedriger Reputation. Nach jeder einer Transaktion folgenden Bewertung wird der systemweit gültige Reputationswert eines Agenten angepasst. Dazu dient folgende Formel (nach Zacharia 1999):

$$R_{t+1} = \frac{1}{\theta}\sum_{1}^{t}\Phi(R_i)R_{i+1}^{other}(W_{i+1} - \frac{R_t}{D})$$

$$\Phi(R) = 1 - \frac{1}{1+e^{\frac{-(R-D)}{\sigma}}}$$

Die neue Reputation (R_{t+1}) entspricht also einer mit dem Gedächtnisfaktor θ (je größer θ, desto größer das Gedächtnis des Systems) gewichteten Summe über alle mit einer Dämpfungsfunktion Φ (σ steuert dabei die Steilheit der Dämpfungsfunktion) modifizierten bisherigen Reputationswerte, die mit der Reputation des diese neue Bewertung abgebenden Agenten (R_{i+1}^{other}) und der eigentlichen Bewertung (W_{i+1}) aus dem Intervall [0.1;1] verrechnet wird. *D* steht dabei für die maximale Größe, die der Reputationswert erreichen kann, bei Zacharia (1999) liegt *D* bei 3000. Wenn W_{i+1} kleiner als die bisherige Reputation geteilt durch die maximale Reputation *D* ist – gedacht als Erwartungswert für die Bewertung – verliert der bewertete Agent an Reputation, sonst gewinnt er dazu. Ziel des Formalismus ist es, dass sich nach einer großen Zahl an Bewertungen der Reputationswert R_{t+1}

[30] Im Zusammenhang mit *Sporas / Histos* eigentlich zu lesen als ‘neue Agenten, bzw. deren EigentümerInnen’, da Zacharia davon ausgeht, dass jeder Agent eine NutzerIn repräsentiert, und nicht völlig autonom handeln kann.

der tatsächlichen Reputation des Agenten annähert (vgl. Moukas et al. 1999: 316). Das System geht also davon aus, dass es tatsächlich so etwas wie eine für alle anderen gültige objektive Reputation eines Agenten gibt.

Da der *Sporas*-Algorithmus mit *Histos* gekoppelt verwendet werden soll, ist anzunehmen, dass die der Bewertung zugrundeliegende Datenstruktur auch bei *Sporas* ein Graph ist, in dem Agenten durch Knoten und die aktuellsten Bewertungen (samt 'Zeitstempel') durch gerichtete Kanten ausgedrückt sind (vgl. auch die Ausführungen zu *Histos*). Die Berechnung der Reputationseinschätzung müsste dann die auf einen Agenten zeigenden Kanten in chronologischer Reihenfolge heranziehen. Wenn ein Multiagenten-System über n Agenten verfügt, die sich – als *worst case* – alle gegenseitig bewertet haben, liegt die Speicherkomplexität einer derartigen Struktur bei $O(n^2)$.

Histos

Wenn ein Agent bereits mit vielen anderen Agenten Kontakt hatte, wird der vom System ermittelte Reputationswert nicht nach der *Sporas*-Formel berechnet, sondern aufgrund des Netzes der paarweisen Bewertungen der NutzerInnen. Histos geht dabei von der Idee aus, dass wir dazu tendieren, der Bewertung eines Freundes eines Freundes eher zu trauen als der Bewertung eines völlig Fremden (Zacharia 1999: 4). Das Netz der paarweisen Bewertungen dazu als ein systemweit existierender gerichteter Graph verstanden, der aus Knoten besteht, die für einzelne Agenten stehen, und aus Kanten, die die jeweils letzte Bewertungen eines Agenten durch einen anderen symbolisieren. Um nun die personalisierte Reputation eines Agenten L zu berechnen (vgl. Zacharia 1999: 4f), untersucht *Histos* diesen Graphen darauf, ob es einen direkten Pfad von A nach A_L gibt. Wenn es einen Pfad der Länge 1 gibt, wurde A_L von A selbst bewertet, und diese Bewertung wird als Reputationswert genutzt. Ansonsten wird eine Breitensuche durchgeführt, um alle Pfade zwischen A und A_L zu finden, wobei diese maximal die Länge N haben dürfen. Für die Berechnung selbst werden nur die θ neusten Pfad bezüglich der Bewertung von A_L zugrundegelegt. Um die Bewertung von A_L durchführen zu können, müssen die personalisierten Reputationswerte der nur einen Schritt von A_L entfernten Agenten bekannt sein. Dazu wird rekursiv wiederum deren Reputation berechnet (Länge dann maximal N-1), bis hin zu den direkt an A grenzenden Agenten, bei denen die Bewertung durch A als Rechengrundlage genutzt werden kann. Die Zeitkomplexität der Rekursion liegt also bei $O(\theta\ N)$. Die jeweilige Berechnung des Reputationswertes erfolgt dabei nach folgender, leicht veränderter Formel (nach Zacharia 1999: 5)

$$R_{t+1} = \frac{1}{\theta^t} \sum_{t-\theta^r}^{t} \Phi(R_{i+1})(R_{i+1}^{other} W_{i+1}) \,/\, \sum_{t-\theta^r}^{t} R_{i+1}$$

$$\theta^t = \min(\theta, m);\ m = \deg(A_L)$$

m steht dabei für die Zahl der Pfade von A nach A_L (wie oben diskutiert). Wenn es einen direkten Pfad gibt, wird die Bewertung dieses Pfades statt R_{t+1} zur Berechnung herangezo-

gen. Für den Fall, dass keine oder nur Pfade mit einer Länge größer als *N* zwischen den beiden Agenten existieren, wird auf den *Sporas*-Mechanismus zurückgegriffen.

Diskussion

In den meisten Fällen scheint der *Sporas / Histos*-Mechanismus die gewünschte 'objektive' Bewertung zu erlauben. Eine Beschränkung des Mechanismus ergibt sich aus der Annahme, dass die Bewertungen nicht von Agenten, sondern von den hinter diesen stehenden NutzerInnen durchgeführt werden. Deswegen – und um einen Mißbrauch des Bewertungssystems zu verhindern – wird ja auch jeweils nur die letzte Bewertung herangezogen. Wenn nun allerdings Agenten statt NutzerInnen Bewertungen aussprechen, kann es zu Problemen kommen. Zum einen erinnert sich das System nur an die jeweils letzte Bewertung. Ein Agent, der abwechselnd gute und schlechte Erfahrungen macht, wird deswegen möglicherweise wieder auf einen betrügerischen Partner zurückgreifen, da die Reputation z.B. nicht unter die eines neuen Agenten sinken kann. Sinnvoll wäre es also, für den Fall, dass Agenten und nicht NutzerInnen die Bewertung vornehmen, Zacharias Vorschlag mit agentenzentrierten Ansätzen zu koppeln. Ein weiteres Problem lässt sich wie folgt konstruieren: Nehmen wir an, ein Agent *B* führt mit jedem anderen Agenten einmal eine Transaktion aus und betrügt dabei nicht. Alle bewerten dementsprechend positiv. In einer zweiten Runde versucht *B* wiederum, mit jedem Agenten ein Geschäft durchzuführen. Diesmal betrügt er jedes Mal und wird auch entsprechend schlecht bewertet. Angenommen, $\theta > 2N$. Da von jedem Agenten aus ein Pfad $< N$ zu *B* gegeben ist, wird bei der Frage nach der Reputation von *B* der *Histos*-Mechanismus herangezogen. Das dort gespeicherte soziale Wissen darüber, dass *B* betrügt, nützt jedem einzelnen Agenten aber nicht, da *Histos* in jedem Fall nur dessen persönliche, gute Erfahrung mit *B* zurückgibt. D.h. in diesem speziellen Fall scheint der Schutz vor Betrug nicht besser als bei Marshs Vorschlag und schlechter als in den Vorschlägen von Schillo oder Yu / Singh zu sein. Der Einsatz eines an Zacharia angelehnten Systems in einer von Winter (1999) durchgeführten Simulation zeigte, dass zentralisierte Reputationsinformationen zu einer Verbesserung der Qualität eines Marktplatzes führen können.

Kriterien

Sporas / Histos kann dazu beitragen, *Online Communities* sicherer zu machen und dort eine Reputationsbewertung bereitzustellen. Die von Zacharia vorgegebenen Kriterien werden erfüllt. Allerdings lässt sich darüber streiten, ob diese Annahmen auch für den Fall eines Multiagenten-Systems ohne NutzerInnen-Feedback die geeignetsten sind, insbesondere über das Kriterium, dass jeweils nur die letzte paarweise Bewertung erinnert wird, und über das Kriterium, dass die Reputation eines Nutzers nie unter den Anfangswert sinken darf. Unklar ist beispielsweise, was dies für Folgen für neu hinzukommende Agenten hat. Die Annahme, dass jeweils nur die letzte paarweise Bewertung hinzugezogen wird, gibt Hinweise auf die technischen Schwierigkeiten eines zentralen Systems, da sowohl die Berechnungskomplexität als auch die Speicherkomplexität deutlich ansteigen würden, wenn nicht nur die letzten θ insgesamt abgegebenen Bewertungen bezüglich eines Agenten,

sondern auch die letzten θ Bewertungen jedes Agenten für jeden anderen gespeichert werden würden. Dementsprechend erscheint es sinnvoll, eine zentrale Reputationsagentur wie den *Sporas / Histos*-Mechanismus mit subjektiven Bewertungsmechanismen mit Gedächtnis in den einzelnen Agenten zu koppeln. Auch aus soziologischer Sicht liegt ein solches Vorgehen nahe, da Zacharia (vgl. 1999) ja von sich selbst sagt, eher eine Institution wie eine Ratingagentur nachbilden zu wollen. Die subjektive Beurteilung der Transaktionen wird den hinter den Agenten stehenden NutzerInnen überlassen. Eine Kopplung mit agentenzentrierten Mechanismen würde diese Subjektivität in die Agenten hineinholen.

4.5 Andere Ansätze

Foner 1999

Nicht richtig in das oben aufgemachte Viererschema hinein passt der von Leonard Foner (1999) vertretene Ansatz, der Ähnlichkeiten mit dem bei *PGP* verwendeten *Web of Trust* (vgl. Foner 1999: 43, Kuhlen 1999b: 358ff) hat. Foner diskutiert im Zusammenhang mit seinem *Matchmaking*-System *Yenta* auch die Frage von Vertrauen, hier vor allem (vgl. auch die Diskussion um *TRUSTe* in Abschnitt 4.3) unter dem Gesichtspunkt des Vertrauens in die Einhaltung von *privacy* (vgl. Foner 1999: 23). Sein Ziel ist es, ein agentenbasiertes System zu entwickeln, dass es ermöglicht, NutzerInnen mit ähnlichen Interessensgebieten zusammenzubringen, und dabei so wenig wie möglich *privacy*-relevante Informationen zu gefährden. Foner beschreibt dazu eine dezentrale Systemarchitektur, die unter Einhaltung hoher Vertraulichkeitsstandards die Zusammenarbeit mehrerer, einander nicht notwendigerweise bekannter Agenten ermöglicht. *Yenta* ist eine Beispielapplikation dafür; die vor allem darauf beruht, dass Empfehlungen ausgehend von lokalen Interessen-Clustern ausgesprochen werden (*collaborative filtering*). Um eine Einschätzung dafür geben zu können, ob ein anderer Nutzer glaubwürdig ist, enthält *Yenta* einen Reputationsmechanismus (vgl. Foner 1999: 43f). Dieser soll aus Selbstbeschreibungen der einzelnen NutzerInnen bestehen, die dem jeweiligen lokalen *Yenta*-Programm bekannt sind und deren Vertrauenswürdigkeit von Dritten – mit einer verschlüsselten digitalen Unterschrift – garantiert wird. Wie bei *Web of Trust* versucht ein Agent bzw. dessen NutzerIn, die Vertrauenswürdigkeit einer unbekannten Person dadurch festzustellen, das überprüft wird, ob eine der die Selbstaussage unterzeichnenden Personen ihm bekannt ist. Ist dies nicht der Fall, wird das selbe Spiel einen Schritt weiter ausgedehnt, und die Vertrauenswürdigkeit der unterzeichnenden Personen ebenso überprüft. Foner vergleicht diesen Mechanismus mit »small-town gossip« (1999: 44), mit dem Unterschied, dass es dort zumeist um Aussagen über Dritte, hier jedoch um Selbstaussagen geht. Ob die entstehende Struktur eher einen netzwerkartigen Charakter hat, oder ob es zur Herausbildung von einzelnen Stellen kommt, die die Selbstaussagen vieler anderer signieren, hängt mit den sozialen und politischen Entscheidungen der NutzerInnenschaft und nicht mit der Systemarchitektur zusammen (vgl. Foner 1999: 44). Im Unterschied zu den meisten anderen hier behandelten Systemen wird Vertrauenswürdigkeit nicht als Erwartungswert über das zukünftige Verhalten oder als

Maß zwischen -1 und +1 realisiert, sondern als explizite Abbildung sozialer Netzwerke; eine Idee, die etwa auch den *Histos*-Mechanismus beeinflusst hat.

Olsson 1998

Eine Erweiterung des von Foner vorgestellten Modells wird von Tomas Olsson (1998) diskutiert. Dazu wird vorgeschlagen, die Ebene des *collaborative filtering* bei *Yenta* um eine Ebene des *social filtering* zu ergänzen. Diese soll aus berechneten Vertrauenswürdigkeitswerten bestehen, die Aussagen darüber machen, wieweit ein Agent einem anderen Agenten in Bezug auf bestimmte Inhalte vertraut. Dieser Vertrauenswürdigkeitswert soll dabei auf durch die NutzerInnen ausgesprochene Bewertungen basieren. Vorschläge würden dann nicht mehr nur aufgrund von Interessenähnlichkeit (verbunden mit der bei Foner diskutierten Signatur, die die Vertrauenswürdigkeit sichern soll) ausgesprochen, sondern in die Auswahl der Vorschläge würde direkt auch die aufgrund der Bewertungen anderer berechnete Vertrauenswürdigkeit einfließen.

Wong / Sycara 1999

H. Chi Wong und Katia Sycara (1999) diskutieren die Frage von Sicherheit und Vertrauenswürdigkeit eines Multiagenten-Systems unter einer etwas anderen Perspektive als die anderen hier behandelten Vorschläge. Um einem MAS Vertrauen hinzuzufügen, schlagen sie verschiedene systemweite Elemente vor. Dazu gehört, vertrauenswürdige *Agent Name Server* und *Matchmaker* zu verwenden, Agenten eineindeutig identifizierbar zu machen und ihnen eine nicht fälschbare Identität zuzuweisen. Agenten müssen – etwa durch Kenntnisse, die nur ihre EigentümerIn besitzt – nachweisen können, dass sie im Auftrag ihrer EigentümerInnen handeln. Damit soll Vertrauen in die Authentizität eines Agenten hergestellt werden. Weiterhin müssen die EigentümerInnen – rechtlich – für Handlung ihres Agenten verantwortlich sein und dafür belangt werden können. Dies soll als Schutz gegen böswillige Agenten dienen. Wong und Sycara führen dann aus, wie mit Hilfe einer *Private / Public-Key*-Infrastruktur diese Vorschläge implementiert werden können. Dieser Ansatz versucht also nicht, gesellschaftliche Mechanismen der Vertrauensbildung auf eine Agentengesellschaft abzubilden, sondern bettet diese in gesellschaftliche und insbesondere rechtlichen Mechanismen ein. Möglicherweise wird dadurch allerdings die Notwendigkeit einer Implementation einer Bewertung der Vertrauenswürdigkeit nur verlagert. Ein tatsächlich autonom handelnder Agent müsste wissen, Agenten welcher EigentümerInnen er lieber meiden sollte. Dazu wären doch Bewertungssysteme sinnvoll.

Winsborough et al. 2000

Winsborough et al. (2000) beschreiben verschiedene Verhandlungsstrategien und ein dafür nutzbares Protokoll, mit dem zwischen *Client* und *Server* verhandelt werden kann. Dieses soll dazu dienen, um automatisierte Verhandlungen über Vertrauenswürdigkeit zwischen einem Agenten und einem potenziellen Handelspartner herzustellen. Den Kern des Verhandlungsprotokoll bildet der Austausch von *property-based digital credentials*, also von Zusicherungen – etwa über die Kreditwürdigkeit, über die Verfügbarkeit einer bestimmten Kreditkarte, über Preise oder auch über die Mitgliedschaft in einer Organisation. Das Grundprinzip der beiden präsentierten Verhandlungsstrategien (*eager negotiation* vs.

parsimonious negotiation) besteht darin, Schritt für Schritt gegenseitig – zunehmend sensiblere – *Credentials* auszutauschen, bis gegenseitiges Vertrauen etabliert ist oder es sich zeigt, dass dieses nicht etabliert werden kann. Von ähnlichen Ansätzen gehen auch Yu et al. (2000) aus.

4.6 Zusammenfassung

Fazit

In diesem Kapitel wurden verschiedene Herangehensweisen an das Problem von Vertrauen bei Multiagenten-Systemen vorgestellt und diskutiert. Dies geschah in den Kategorien agentenzentrierter solitärer bzw. sozialer Ansätze einerseits und externer Ansätze unter Zuziehung ‘objektive’ bzw. ‘subjektiver’ Informationen andererseits. Die Einteilung hat sich bei der Diskussion als hilfreich erwiesen. Neben den durch dieses Raster abgedeckten Ansätzen wurden weitere Vorschläge andiskutiert. Die Ansätze von Marsh (1994b), Schillo (1999) und Zacharia (1999) wurden stellvertretend für solitäre, soziale und externe Modelle ausführlicher vorgestellt und diskutiert. Insgesamt zeigte sich, dass alle Ansätze einzelne Schwachpunkte bezüglich der Systemsicherheit aufwiesen. Ein kombinierter Zugriff auf mehrere Formalismen scheint deswegen empfehlenswert, also etwa die Kombination eines agentenzentrierten, sozialen Ansatzes mit spezialisierten Bewertungsagenturen. Nicht alle vorgestellten Ansätze gingen von soziologischen Modellen des Vertrauens aus; die, die als Vorbild gesellschaftliche Mechanismen genommen haben, machten bei der Umsetzung meist deutliche Einschränkungen in den Randbedingungen. Die Komplexität der verschiedenen Ansätze hängt vor allem davon ab, wie viele Informationen – je Agent bzw. systemweit – über zurückliegende Interaktionen und daraus resultierende Bewertungen gespeichert werden müssen. Dieser Aufwand ist ebenso wie der Rechenaufwand in sozialen Modellen naturgemäß größer als in solitären Ansätzen und begrenzt bei zentralen Systemen die Skalierbarkeit eines MAS.

5 Hinweise zur Implementierung in AVALANCHE

5.1 Beschreibung des in AVALANCHE verwendeten Modells

Padovan et al. 2000

Das in *Java* implementierte Multiagenten-System AVALANCHE wird als Prototyp eines elektronischen Marktplatzes angesehen (vgl. Abschnitt 3.3). AVALANCHE nutzt ein agentenzentriertes, solitäres Reputationsmodell, allerdings mit Anleihen an ein externes Modell. Diese Anleihen an ein externes Modell ergeben sich daraus, dass AVALANCHE von domänenspezifischen *Rating*-Agenten (oder -Agenturen) ausgeht, die jeweils spezifische Reputationsinformationen bereitstellen können. Die einzelnen Agenten in AVALANCHE haben das Ziel, Güter auf einem von mehreren miteinander vernetzen Handelsplätzen zu handeln. Als Teil ihrer Verhandlungsstrategie verwenden sie dabei einen agentenspezifischen Reputationskoeffizienten, der zwischen 0 (schlechter Ruf) und 1 (guter Ruf) schwanken kann, für jeden ihnen bekannten anderen Agenten. Dieser wird dazu herangezogen, um mögliche Transaktionspartner nicht nur aufgrund des Preises, sondern

auch unter Einbeziehung des zu erwartenden Verlustes zu vergleichen. Dazu wird der Reputationskoeffizient als die Wahrscheinlichkeit für das Nicht-Eintreten des Verlustes gewertet, so dass bei einer hundertprozentigen Annahme eines Verlustes beispielsweise der doppelte Preis als Verhandlungsgrundlage genommen wird. (Padovan et al. 2000: 9ff)

Veränderung Reputation

Nach einer Transaktion ändern die beteiligten Agenten ihre gegenseitige Einschätzung. Eine erfolgreiche Transaktion wird mit dem Wert r=1 bewertet, eine fehlgeschlagene mit dem Wert r=0. Die Berechnung eines neuen Reputationskoeffizienten für den Agenten Y aus Sicht von X erfolgt dabei unter Einbeziehung eines Gewichtungswertes α nach folgender Formel (vgl. Abb. 6 für die Veränderung der Reputation nach dieser Formel für r stetig = 1, r stetig = 0, r abwechselnd 0 und 1; α=0,15; R_0=0,5):

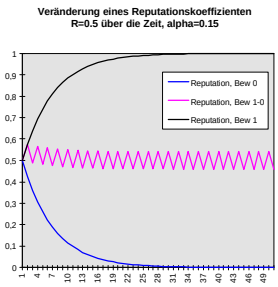


Abb. 6 – Veränderung der Reputation

$$R^{X}_{Y\ t+1} = R^{X}_{Y\ t}(1-\alpha)+r_t\alpha$$

Eine wichtige Eigenschaft dieses Modells ist die Tatsache, dass kooperatives und nicht-kooperatives Vertrauen beides mit dem Maß α (bzw. -α) belegt wird. Die soziologische Konzeptualisierung von Vertrauen legt nahe, Betrug bzw. das Nicht-Zustandekommen einer Kooperation stärker mit Reputationsverlust zu bestrafen (vgl. Kapitel 2), wie dies beispielsweise im Modell von Yu / Singh (2000) mit einem von α (Einbeziehung von Kooperation) unterschiedenen, größeren Wert β (Einbeziehung von Betrug) vorgenommen wurde.

Unbekannte Agenten

Zur Berechnung des Vertrauens in unbekannte Agenten schlagen Padovan et al. (2000) vor, entweder den Mittelwert über alle bekannten Reputationsbewertungen aller Agenten zu nehmen, oder, so dieser nicht verfügbar ist, einen systemspezifischen Defaultwert. Existiert eine möglicherweise domänenspezifische *Rating*-Agentur (also ein spezieller, für die Bewertung anderer Agenten zuständiger Agent), dann kann auch diese um Rat gefragt werden.

***Rating*-Agentur**

Die Einführung einer oder mehrerer *Rating*-Agenturen wird als Erweiterung des derzeitigen AVALANCHE-Modells von (Padovan et al. 2000: 10) diskutiert. Ähnlich wie bei Zacharia (1999) wären diese Agenturen ein ‘subjektives’ externes Element. Einzelne Agenten können (oder müssen) ihre Bewertungen an die Agentur weiterleiten. Diese speichert für jeden Agenten X einen Reputationswert R_X, der ähnlich wie die agentenspezifischen Reputationskoeffizienten durch die Bewertungen der einzelnen Agenten verändert wird. Die einzelnen Bewertungen werden nicht gespeichert, d.h. anders als bei Zacharia kann ein Agent auch ständig Bewertungen über einen anderen Agenten abgeben und damit auch tatsächlich die Reputation beeinflussen. Zwei Agenten könnten so gegenseitig ihren Koeffizienten bei der *Rating*-Agentur hochtreiben. Berechnet werden die Veränderungen des agentenspezifischen, zentralen Reputationswertes R_{Yt+1} nach Bewertung durch X (r_t^X)

wie folgt (Padovan et al. 2000: 11)[31]:

$$R_{Y_{t+1}} = R_{Y_t}(1-\beta) + r_t^X \beta;\ \text{mit}\ \beta = \gamma R_{X_t}$$

γ ist dabei ein agenturspezifisches Äquivalent zu α in der Formel für die Veränderung der Reputation bei den einzelnen Agenten. Sind noch keinerlei Informationen über den Agenten *Y* in der Agentur bekannt, wird statt R_{Yt} der Mittelwert über alle bekannten Reputationen eingesetzt.

Zusammenfassung

Ein Agent in AVALANCHE hat also zum einen die Möglichkeit, seine eigenen Erfahrungen für die Bewertung eines potenziellen Kooperationspartners heranzuziehen, und zum anderen kann er auf den von einer zentralen Agentur kombinierten Erfahrungswert zurückgreifen, der allerdings nicht 'fälschungssicher' ist, d.h. zwei Agenten können theoretisch über häufige gegenseitige positive Bewertungen schnell ihre Reputation bei einer *Rating*-Agentur hochtreiben. Eine direkte soziale Komponente (etwa mit *TrustNet* vergleichbar) gibt es nicht. Bei der Veränderung der Reputationskoeffizienten haben positive und negative Erfahrungen das gleiche Gewicht. Da Agenten bisher nur jeweils ein Produkt kaufen bzw. verkaufen, stellt sich die Frage eines situationsspezifischen Vertrauens bisher nicht. Nach den am Ende von Kapitel 3 genannten Kriterien erfüllt die Einbeziehung eines Reputationskoeffizienten den gewünschten Schutz vor Betrug noch nicht in allen Fällen. Technisch gesehen ist das gewählte Modell effizient, da je anderem Agent nur jeweils ein Wert gespeichert bzw. abgerufen werden muss.

5.2 Diskussion und Empfehlungen

Quantitativer Ansatz

Anders als Marsh (1994a) und Yu / Singh (2000), die den Wertebereich [-1; +1] vorgeschlagen haben, und wie Schillo (1999) und auch Gambetta (1988b) verwendet AVALANCHE den Wertebereich [0; 1] für den Reputationskoeffizienten und geht von Wahrscheinlichkeitsrechnungen aus. Dafür spricht einiges, nicht zuletzt eine relativ einfache Handhabung der verwendeten Mathematik und die Überlegungen von Gambetta (1988b) zur Konzeption der Reputation bzw. Vertrauenswürdigkeit in Bezug auf die Wahrscheinlichkeit kooperativen bzw. nicht kooperativen Verhaltens. Die Nutzung des Bereichs [-1; 1], obwohl auf den ersten Blick plausibel erscheinend, führt bei Marsh zu problematischen Ergebnissen und bei Yu / Singh zu einer relativ komplizierten und viele Sonderfälle berücksichtigenden Formalisierung (vgl. Tabelle 7). Andere hier vorgestellte Ansätze gehen von einem Bereich [0; 3000] aus (Zacharia 1999, ohne dies allerdings zu begründen), sehen ein unbeschränktes Wachstum der Reputation sowohl nach unten als auch nach oben vor (*eBay*), oder stellen eher qualitative Formalisierungen in den Vorder-

[31] Theoretisch müssten bei zwei sich gegenseitig bewertenden Agenten beide Bewertungen gleichzeitig von der Agentur bearbeitet werden, um Verfälschungen zu vermeiden.

grund (vgl. vor allem Abdul-Rahman / Halles 1997). Aus Sicht der Soziologie bleibt es fraglich, ob sich Vertrauenswürdigkeit einfach, möglicherweise sogar linear, quantifizieren lässt. Zwischen Vertrauen und Misstrauen liegen ebenso wie zwischen einzelnen Stufen der Vertrauenswürdigkeit große Unterschiede (vgl. vor allem Luhmann 1989). Gambetta (1988b) kann so interpretiert werden, dass das Wachsen und Sinken der Vertrauenswürdigkeit je nach erreichtem Vertrauenslevel und je nach vorherigem Vertrauen oder Misstrauen in unterschiedlicher Geschwindigkeit und mit unterschiedlichem Gewicht erfolgt; dies wird in der AVALANCHE zugrundeliegenden Formalisierung ebenfalls nicht berücksichtigt.

Reputation < 0.5

Es werden in (Padovan et al. 2000) keine Aussagen darüber gemacht, was geschehen soll, wenn ein Agent eine Reputation unterhalb von 0,5 hat, also wahrscheinlich häufiger betrügt, als das er nicht betrügt. Unter der Annahme, dass der Preis mit der erwarteten Verlustwahrscheinlichkeit verrechnet wird, könnte sich ein betrügerischer Agent – je nach Wahl von α, und je nachdem, ob nur persönliche Erfahrungen oder auch generelle Informationen zählen – nahezu beliebig oft wieder an Geschäften mit demselben Agenten beteiligen. Er muss dazu nur einen Preis anbieten, der ausreichend niedrig ist.[32] Da das Produkt eh nicht geliefert wird, hat der betrügende Agent dadurch auch keine Kosten. Um dies zu umgehen, müssten Agenten entweder bei Preisen, die deutlich unter dem Durchschnitt des Handelsplatz liegen, skeptisch werden, oder die einfache *Ranking*-Funktion, die alle Angebote nach dem um den wahrscheinlichen Verlust erhöhten Preis ordnet, müsste Agenten mit einer Reputation unterhalb eines agentenspezifischen, vielleicht auch von der Wichtigkeit eines Geschäftes abhängigen Schwellenwertes (wie dies Marsh 1994a vorschlägt) ganz ausschließen. Damit stellt sich allerdings wieder die Frage nach der Reputation neuer Agenten – diese dürfte dann jedenfalls nicht 0 betragen. Eine andere Möglichkeit bestände darin, die Verlustwahrscheinlichkeit mit einer Potenz- oder Exponentialfunktion so zu gewichten und dies mit dem Ausschluss einer Verlustwahrscheinlichkeit > 0.99 zu verbinden.

Informationsökonomie

Generell erscheint es wünschenswert, den Austausch und die Anforderung von Informationen in das Verhandlungsprotokoll der AVALANCHE-Agenten mit aufzunehmen. Dies bezieht sich nicht nur auf die Möglichkeit, Reputationsinformationen von einer *Rating*-Agentur (und vielleicht sogar von 'ganz normalen' anderen Agenten) anzufordern und an diese weiterzugeben. Eventuell lassen sich damit sogar Marktnischen finden. Auch die Möglichkeit, von einem Agenten vor dem Abschluss einer Transaktion Informationen zur Authentifizierung seiner Identität, zur Identität der EigentümerIn und zu einem möglichen Zahlungsrahmen anzufordern, gehört zu den möglichen informationellen Geschäften zwischen Agenten – auch im Sinne einer Herausbildung von *brand names*. Damit müsste

---

[32] Vgl. zu diesem Dilemma des Betrugs bei bekanntem Algorithmus auch Marsh (1994a: 138).

zwar einerseits eine entsprechende Ontologie geschaffen oder genutzt werden, zum anderen würde Reputation aber nicht mehr nur auf eine Wahrscheinlichkeitskennziffer beschränkt bleiben, sondern könnte, wie im Abschnitt 2.2 diskutiert, auf andere Hinweissysteme wie die Mitgliedschaft der EigentümerIn in einer vertrauenswürdigen Organisation oder eine Zusicherung einer Bank über einen ausreichenden Kreditrahmen für den Kauf eines Produktes ausgedehnt werden. Anregungen dazu geben die in Abschnitt 4.5 diskutierten AutorInnen. Mit der Einbeziehung derartiger Faktoren in die Entscheidung eines Agenten müsste dieser allerdings auch deutlich an Komplexität zunehmen und möglicherweise auf Methoden der KI-Forschung zurückgreifen. Zu diskutieren (dies geschieht am Ende dieses Abschnitts) wäre auf jeden Fall auch die Frage, wie autonom ein Agent sein soll und wieweit Eingriffe der EigentümerIn notwendig sind.

***Rating*-Agentur**

In Bezug auf die angedachte *Rating*-Agentur bleibt unklar, wie diese in das Gesamtgefüge von AVALANCHE eingebunden werden soll. Soll es eine Agentur je Handelsplatz geben, an die dann auch alle Bewertungsinformationen – als Teil eines Agentenstandards – geliefert werden müssen? Soll es gar eine systemweite Agentur geben? Oder sollen mehrere, möglicherweise miteinander konkurrierende Agenturen – auch zur selben Domäne und auf dem gleichen Handelsplatz – existieren? An welche Agentur wenden Agenten sich dann, von welcher erhalten sie Ratschläge? Unklar bleibt auch, ob es die Agenten etwas kosten soll, Informationen von der *Rating*-Agentur zu erhalten oder Informationen an diese zu liefern (oder ob die Agentur – im Fall der Informationslieferung – gar dafür bezahlt). Ist die Agentur nicht mit dem Handelsplatz identisch (und kann so z.B. bei der Annahme von Bewertungen auch prüfen können, ob überhaupt ein Transaktionsversuch vorlag), müsste auf jeden Fall noch eine Lösung für das Problem des gegenseitigen Aufschaukelns von Reputationen durch zwei zusammenarbeitende Agenten gefunden werden. Zacharia (1999) schlägt für dieses Problem vor, nur die jeweils letzten Informationen zu berücksichtigen – dann würde allerdings die *Rating*-Agentur erheblich komplexer, da sie speichern müsste, welche Agenten schon einmal was über welche anderen Agenten berichtet haben. Auch der Reputationswert je Agent müsste dann allerdings bei jeder erneuten Bewertung komplett neu berechnet werden.

Soziale Elemente

Die Verknüpfung von der Bewertung eigener Erfahrungen mit einer *Rating*-Agentur erscheint als relativ flexible und einfach umsetzbare Lösung. Bisher nicht betrachtet wurde die Frage, ob ein agentenzentrierter, sozialer Ansatz eine brauchbarere Lösung als ein solitärer Ansatz darstellt. Insbesondere mit Schillo (1999) und Yu / Singh (2000) gibt es ja durchaus Ideen, wie ein derartiger Ansatz umgesetzt werden könnte; bleibt es bei *Rating*-Agenturen, könnte hier auf den *Histos*-Mechanismus (Zacharia 1999) zurückgegriffen werden, um ähnliche Effekte zu erzielen. Der Vorteil einer soziales Ansatzes liegt darin, dass – sollte es einen *Gossip*-Mechanismus geben – Informationen über betrügerische Agenten sehr schnell verbreitet werden könnten, was den Handelsplatz insgesamt besser

vor böswilligen Agenten schützt. Der Nachteil liegt in einer steigenden Komplexität entweder aller Agenten oder zumindest der *Rating*-Agentur. Neben einer *TrustNet*-ähnlichen Datenstruktur für die Speicherung der einzelnen Bewertungen müssten Agenten in einem solchen Fall auch ein über *Kaufen / Verkaufen* hinausgehendes Kommunikationsprotokoll besitzen. Dies wirkt sich insbesondere bei einem angedachten offenen System stark auf die Komplexität der ja aus ganz unterschiedlichen Quellen kommen könnenden Agenten aus. Ein weiteres Problem eines sozialen Mechanismus liegt darin, dass diese zumeist Nachbarschaften irgendeiner Art heranziehen, um Agenten zu finden, die möglicherweise Informationen über einen anderen Agenten besitzen könnten, und so zu Bewertungspfaden gelangen. Da nicht jeder Agent alle anderen Agenten kennen kann (und es z.B. eine Funktion des Systems ist, einem Agenten mögliche passende Handelspartner zu nennen), stellt sich hier die Frage, was in AVALANCHE eine Nachbarschaft sein könnte. In vielen der Simulationen (z.B. Marsh 1994a; Rasmusson 1996) wurde dafür die räumliche Nähe in einem imaginären Raum verwendet. Die AVALANCHE-Agenten verfügen (außer durch die Wahl eines Handelsplatzes) nicht über räumliche Koordinaten; dies erscheint auch nicht sinnvoll. Sofern nicht sämtliche Agenten eines Handelsplatzes als eine Nachbarschaft zählen sollen, könnte dies z.B. auf vertrauenswürdig erscheinende, einem Agenten bekannte Agenten auf dem selben Handelsplatz eingeschränkt werden – diese wiederum müssten dann allerdings auch ‘angefunkt’ werden können, wobei sich wiederum die Frage stellt, in welchen Phasen des Handelsprotokolls das möglich sein soll. Festhalten lässt sich jedenfalls, dass die Einführung eines sozialen Modells AVALANCHE deutlich komplizieren würde. Selbst wenn dies – vorerst – nicht geschehen soll, bleibt die grundlegende Frage, wie die Kommunikationsprotokolle der Agenten verändert werden müssten, um auf eine derartige Situation zumindest mit ignorieren reagieren zu können.

Stereotypen

Eine in keinem der diskutierten Systeme eingeführte, aber durchaus mögliche Erweiterung der komplexitätsreduzierenden Funktion von Vertrauen / Reputation läge in der Einführung einer Möglichkeit, dass ein Agent ausgehend von bestimmten Ähnlichkeitsmerkmalen Stereotypen über andere Agenten (oder quasi Cluster von Agenten) bildet, und sich bei neuen Agenten, die dann auch ähnliche Merkmale haben, von der durchschnittlichen Reputation dieser Cluster leiten lässt. Allerdings ergibt ein derartiger, durchaus von der Soziologie gedeckter Ansatz (vgl. Misztal 1996) in einem Multiagenten-System nur dann Sinn, wenn sowohl Unterscheidungsmerkmale als auch ein relativ ähnliches Verhalten ähnlicher Agenten gegeben ist.

Sicherheitsarchitektur

In dieser Arbeit nur am Rande angesprochen wurde die eine wichtige Bedeutung besitzende Sicherheitsarchitektur des Gesamtsystems. Insbesondere die in Abschnitt 4.5 diskutierten Ansätze machen umfangreiche Vorschläge, wie sowohl ein Schutz von *privacy* als auch ein technischer Schutz vor Missbrauch mit Hilfe kryptographischer Techniken und digitaler Signaturen erreicht werden kann. Diese Problemstellungen sollten auf dem Weg vom

Prototyp zum tatsächlichen Produkt auch bei AVALANCHE unbedingt berücksichtigt werden. Rasmusson (1996) ebenso wie Kuhlen (1999a; 1999b) und Foner (1999) betonen die Gefahr, die die Konzentration auf eine zentralisierte Sicherheitseinrichtung darstellt, etwa auf eine zentrale 'Schlüsselverwahranstalt' beim Einsatz von Verschlüsselungstechniken. Foner (1999) zeigt einige Wege auf, wie ein sicheres System ohne den Einsatz zentraler Server auskommt und ohne die Notwendigkeit der Identifizierbarkeit auskommt. Wong / Sycara (1999) stellen hingegen (ebenso wie Rasmusson 1996) Überlegungen an, wie mit Hilfe von Authentifizierungstechniken die vor allem auch rechtliche Verantwortlichkeit der EigentümerInnen sichergestellt werden kann. Auf AVALANCHE bezogen bietet sich – insbesondere aus rechtlicher Sicht – die Bereitstellung sicherer Übertragungskanäle kombiniert mit den diskutierten Elementen der Sicherheit durch Vertrauen / Reputation kombiniert mit einer eineindeutigen Identifizierbarkeit der beteiligten Agenten und ihrer EigentümerInnen an.

Autonomie der Agenten

Damit stellt sich auch schon die abschließende Frage, welche Rolle die NutzerInnen bzw. EigentümerInnen der Agenten in AVALANCHE erhalten sollen (vgl. Kuhlen 1999a, 1999b; Maes et al. 1998). Je transparenter und kontrollierbarer die Aktionen ihrer Agenten aus Sicht der EigentümerInnen sind, desto stärker ist das Vertrauen in das Gesamtsystem. Dies bezieht sich nicht nur auf die Einstellung von Faktoren wie der Verhandlungsstrategie oder monetären Grenzwerten, sondern auch auf die Frage, ob vor endgültigem Abschluss einer Transaktion beispielsweise die Zustimmung der jeweiligen EigentümerInnen vorliegen muss, die ja beispielsweise nicht durch den Agentenmarktplatz abbildbare Bedenken gegen bestimmte Handelspartner haben könnten. Eine weitere Möglichkeit könnte darin bestehen, dass Agenten in Zweifelsfällen bei ihren EigentümerInnen nachfragen. Auch die Wahl des Handelsplatzes und schließlich die Frage der Bewertung sind zu diskutieren. Soll – wie in *Kasbah* (vgl. Zacharia 1999; Maes et al 1998) die Bewertung der Transaktion durch die Agenten erfolgen, oder sollen auch hier die End-NutzerInnen hinzugezogen werden? Soll schließlich die Transaktion selbst durch die Agenten ausgelöst werden, oder kommt es an dieser Stelle zu menschlichen Eingriffen? Je nach Anwendungsgebiet und Art der angebotenen Waren sind die hier zu treffenden Entscheidungen sicherlich unterschiedlich. In allen Fällen sollten die Agenten jedenfalls ihre EigentümerInnen ausführlich über die getätigten Transaktionen und die Konditionen des ausgewählten Internet-Marktplatzes informieren. Eng mit den hier genannten Fragen hängt schließlich noch die Frage zusammen, wieweit Agenten standardisiert, vom Marktplatz oder spezialisierten Dritten angeboten oder prinzipiell von jeder NutzerIn programmiert werden können. Auch dabei entstehen jeweils unterschiedliche Problemlagen.

5.3 Zusammenfassung

Fazit

In diesem Kapitel wurde das System AVALANCHE vorgestellt. Ausgehend von der Diskussion über Vertrauen in den Kapitel 2 (Soziologische Grundlagen), 3

(Übertragbarkeit auf technische Systeme) und 4 (Mögliche technische Lösungen) wurden verschiedene Elemente des Reputationssystems in AVALANCHE diskutiert. Um die Sicherheit des Gesamtsystems zu erhöhen, wurde vorgeschlagen, den Informationsaustausch zwischen den Agenten zu erweitern, sowohl im Hinblick auf evtl. soziale Komponenten als auch im Hinblick einer Einbeziehung von Identitäts-Zertifikaten u.ä. in die Verhandlungsstrategien. Betont wurde die Bedeutung einer Sicherheitsarchitektur und die Frage der Verantwortlichkeit der EigentümerInnen der Agenten; beides Aspekte, die nicht nur zu Vertrauen zwischen den Agenten, sondern auch zum Vertrauen in das Multiagenten-System beitragen.

6 Literatur

ABDUL-RAHMAN / HALLES 1997
Alfarez Abdul-Rahman and Stephan Halles: »A Distributed Trust Model«, in *Proceedings of the workshop on new security paradigms*, ACM 1997, pp. 48-60.

AXELROD 1991
Robert Axelrod: *Die Evolution der Kooperation*. 2. Aufl., München: Oldenbourg 1991.

BACHMANN 1998
Reinhard Bachmann: »Kooperation, Vertrauen und Macht in Systemen Verteilter Künstlicher Intelligenz. Eine Vorstudie zum Verhältnis von soziologischer Theorie und technischer Modellierung«, in Thomas Malsch (Hrsg.): *Sozionik.* Berlin: edition sigma 1998, S. 197-234.

BRAUN 1998
Holger Braun: »The Role-Taking of Technology. Vom Sozialwerden der Technik«, in Thomas Malsch (Hrsg.): *Sozionik.* Berlin: edition sigma 1998, S. 169-195.

CASTELFRANCHI ET AL. 1997
C. Castelfranchi, F. de Rosis and R. Falcone: »Social Attitudes and Personalities in Agents«, in K. Dautenhahn et al.: *Socially Intelligent Agents. Paper from the 1997 AAAI Fall Symposium, November 8-10, Cambridge, Mass.*, Technical Report FS-97-02, 1997 (zitiert nach Schillo 1999).

DASGUPTA 1988
Partha Dasgupta: »Trust as a Commodity«, in Diego Gambetta (ed.): *Trust. Making and Breaking Cooperative Relations.* New York / Oxford: Basil Blackwell 1988, pp. 49-72.

DEUTSCH 1973
Morton Deutsch: *The Resolution of Conflict. Constructive and Destructive Processes.* New Haven and London: Yale University Press 1973, S. 143-214.

EYMANN ET AL. 1998
Torsten Eymann, Detlef Schoder, Boris Padovan: »The Living Value Chain. Coordinating Business Processes with Artificial Life Agents«, in Nwana, Hyacinth S., Ndumu, Divine T. (Eds.): *Proceedings of the 3rd Intl. Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM'98)*, Blackpool: The Practical Application Company 1998, p. 111-122.

EYMANN / PADOVAN 1999
Torsten Eymann, Boris Padovan: »Eine Multiagenten-Simulation zur ökonomischen Analyse der dezentralen Koordination von Wertschöpfungsketten«, in August-Wilhelm Scheer; Markus Nüttgens (Hrsg.): *Electronic Business Engineering / 4. Internationale Tagung Wirtschaftsinformatik* 1999. Heidelberg: Physica-Verlag 1999, S. 625-641.

FONER 1999
Leonard N. Foner: *Political Artifacts and Personal Privacy: The Yenta Multi-Agent Distributed Matchmaking System.* PhD thesis, Cambridge: MIT 1999.

GAMBETTA 1988a
Diego Gambetta (ed.): *Trust. Making and Breaking Cooperative Relations.* New York / Oxford: Basil Blackwell 1988.

GAMBETTA 1988b
Diego Gambetta: »Can We Trust Trust?«, in Diego Gambetta (ed.): *Trust. Making and Breaking Cooperative Relations.* New York / Oxford: Basil Blackwell 1988, pp. 213-237.

GILBERT / TROITZSCH 1999
Nigel Gilbert; Klaus G. Troitzsch: *Simulation for the Social Science*. Open University Press 1999.

GOOD 1988
David Good: »Individuals, Interpersonal Relations, and Trust«, in Diego Gambetta (ed.): *Trust. Making and breaking Cooperative Relations.* New York / Oxford: Basil Blackwell 1988, pp. 31-48.

JONES / MARSH 1997
Steve Jones and Steve Marsh: »Human-Computer-Human Interaction: Trust in CSCW«, in *SIGCHI*

Bulletin Vol. 29, No. 3, July 1997. [http://www.cwi.nl/~steven/sigchi/bulletin/1997.3/ jones.html]

JUNGE 1998
Kay Junge: »Vertrauen und die Grundlagen der Sozialtheorie. Ein Kommentar zu James S. Coleman«, in H.-P. Müller, M. Schmid (Hrsg.): *Norm, Herrschaft und Vertrauen: Beiträge zu James S. Colemans Grundlagen der Sozialtheorie.* Opladen: Westdeutscher Verlag 1998, S. 26-63.

KOLLOCK 1994
Peter Kollock: »The Emergence of Exchange Structures: An Experimental Study of Uncertainty, Commitment, and Trust«, *in American Journal of Sociology*, Vol. 100, No. 2, 1994, pp. 313-345.

KUHLEN 1999a
Rainer Kuhlen: »Vertrauen und konstruktives Mißtrauen auf elektronischen Märkten«. Vortrag im Kolloquium »EDV und Recht« der Universität Konstanz, 27. April 1999; [http://www.inf-wiss.uni-konstanz.de/People/RK/Texte/jura_kn99.pdf].

KUHLEN 1999b
Rainer Kuhlen: *Die Konsequenzen von Informationsassistenten. Was bedeutet informationelle Autonomie oder wie kann Vertrauen in elektronische Dienste in offenen Informationsmärkten gesichert werden?* Frankfurt am Main: Suhrkamp 1999.

LUHMANN 1989
Niklas Luhmann: *Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität.* 3., durchgesehene Auflage. Stuttgart: Ferdinand Enke Verlag 1989.

LUHMANN 1998
Niklas Luhmann: *Die Gesellschaft der Gesellschaft.* Bd. 1, Frankfurt am Main: Suhrkamp 1998.

MAES ET AL. 1998
Pattie Maes, Robert H. Guttman, Alexandros G. Moukas: *Agents that Buy and Sell: Transforming Commerce as we Know it.* MIT Media Lab, 1998. [http://ecommerce.media.mit.edu/papers/cacm98.pdf]

MALSCH 1997
Thomas Malsch: »Die Provokation der 'Artificial Societies'. Ein programmatischer Versuch über die Frage, warum die Soziologie sich mit den Sozialmetaphern der Verteilten Künstlichen Intelligenz beschäftigen sollte«, in *Zeitschrift für Soziologie*, Jg. 26, Heft 1, Februar 1997, S. 3-21.

MALSCH ET AL. 1998
Thomas Malsch, Michael Florian, Michael Jonas, Ingo Schulz-Schaeffer: »Sozionik. Expeditionen ins Grenzgebiet zwischen Soziologie und künstlicher Intelligenz«, in Thomas Malsch (Hrsg.): *Sozionik*, Berlin: edition sigma 1998, S. 9-24.

MANHART 1999
Klaus Manhart: »Künstlich sozial. Die Informatik entdeckt die Soziologie«, in *c't* 21/1999, S. 134-140.

MARSH 1992
Stephen Marsh: »Trust and Reliance in Multi-Agent Systems: A Preliminary Report«, in *MAAMAW '92, 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, Rome 1992.

MARSH 1994a
Stephen Paul Marsh: *Formalising Trust as a Computational Concept.* PhD Thesis, Dept. of Computing Science and Mathematics, University of Stirling 1994. [über http://ai.iit.nrc.ca/~steve/Publications.html].

MARSH 1994b
Stephen Marsh: *Optimism and pessimism in trust.* [http://ai.iit.nrc.ca/~steve/Publications.html] 1994.

MISZTAL 1996
Barbara A. Misztal: *Trust in Modern Societies.* Cambridge: Polity Press 1996.

MOUKAS ET AL. 1999
Alexandros Moukas, Giorgos Zacharia and Pattie Maes: »Amalthaea and Histos: MultiAgent Systems for WWW Sites and Reputation Recommendations«, in M. Klusch (ed.): *Intelligent Information Agents. Agent-Based Information Discovery and Management on the Internet.* Berlin / Heidelberg / New York: Springer 1999, pp. 292-322.

MUELLER 1995
Frank Mueller: »Organizational Governance and Employee Cooperation: Can We Learn from Economists?«, in *Human Relations*, Vol. 48, No. 10, 1995, pp. 1217-1235.

OLSSON 1998
Tomas Olsson: »Decentralised social filtering based on trust«, in *AAAI-98 Recommender Systems Work Papers*, 1998, pp. 84-88. [über http://www.sics.se/isl/ publications.shtml].

PADOVAN ET AL. 2000
Boris Padovan, Stefan Sackmann, Torsten Eymann: »Secure Electronic Marketplaces based on the multi agent system AVALANCHE«. Accepted paper for the *11th International Conference on Information and Intelligent Systems (IIS 2000)*, September 20th – 22nd, Varazdin, Croatia..

PREISENDÖRFER 1995
Peter Preisendörfer: »Vertrauen als soziologische Kategorie. Möglichkeiten und Grenzen einer entscheidungstheoretischen Fundierung des Vertrauenskonzepts«, in *Zeitschrift für Soziologie*, 4 (24), 4. August 1995, S. 263-272.

RASMUSSON 1996
Lars Rasmusson: *Socially Controlled Global Agent Systems.* Master's Thesis, Royal Institute of Technology, Dept. of Computer and Systems Science, Stockholm 1996. [über http://www.sics.se/isl/publications.html].

RASMUSSON ET AL. 1997
Lars Rasmusson, Andreas Rasmusson, Sverker Janson: »Using Agents to Secure the Internet Marketplace. Reactive Security and Social Control«, in *Proceedings of Practical Applications of Agents as Multi-Agent Systems 1997 (PAAM'97).* London 1997. [über http://www.sics.se/isl/publications.shtml].

RASMUSSON / JANSON 1996
Lars Rasmusson and Sverker Janson: »Simulated social control for secure Internet commerce«, in *New Security Paradigms '96. ACM Press,* 1996. [über http://www.sics.se/isl/publications.html].

RIPPERGER 1998
Tanja Ripperger: *Ökonomik des Vertrauens: Analyse eines Organisationsprinzips.* Tübingen: Mohr Siebeck 1998.

SACKMANN 1998
Stefan Sackmann: Modellierung ökonomischen Verhaltens in Multi-Agenten-Systemen. Diplomarbeit, Albert-Ludwigs-Universität Freiburg i.Br. 1998. [http://omnibus.uni-freiburg.de/~sackmann/mas/inhalt.htm].

SCHILLO 1999
Michael Schillo: *Vertrauen und Betrug in Multi-Agenten-Systemen. Erweiterung des Vertrauensmodells von Castelfranchi und Falcone um eine Kommunikationskomponente.* Dipl.-Arbeit, Universität der Saarlandes 1999. [über http://www.virtosphere.de/schillo/research/index.html]

SCHILLO ET AL. 1999
Michael Schillo, Petra Funk, Michael Rovatsos: »Who can you Trust: Dealing with Deception«, in C. Castelfranchi, Y. Tan, R. Falcone and B. S. Firozabadi (eds.), *Proceedings of the Workshop »Deception, Fraud and Trust in Agent Societies« of the Autonomous Agents Conference,* 1999.

SCHULZ-SCHAEFFER 1998
Ingo Schulz-Schaeffer: »Akteure, Aktanten und Agenten. Konstruktive und rekonstruktive Bemühungen um die Handlungsfähigkeit von Technik«, in Thomas Malsch (Hrsg.): *Sozionik.* Berlin: edition sigma 1998, S. 129-167.

SZTOMPKA 1999
Piotr Sztompka: *Trust. A Sociological Theory.* Cambridge: University Press 1999.

THIMBLEBY ET AL. 1994
Thimbleby, S. Marsh, S. Jones, A. Cockburn: »Trust in CSCW«, in S. Scrivener (ed.): *Computer-Supported Cooperative Work.* Aldershot: Avebury Technical 1994. (zitiert nach Bachmann 1998 bzw. nach Marsh 1994a)

WINSBOROUGH ET AL. 2000
W. Winsborough, K. Seamons, and V. Jones: »Automated Trust Negotiation«, submitted for journal publication, April 2000. [http://drl.cs.uiuc.edu/security/pubs.html].

WINTER 1999
Mike Winter: »The Role of Trust and Security Mechanisms in an Agent-Based Peer Help System«, in C. Castelfranchi et al. (eds.): *Deception, Fraud and Trust in Agent Societies,* Seattle 1999, pp. 139-148.

WONG / SYCARA 1999
H. Chi Wong and Katia Sycara: »Adding Security and Trust to Multi-Agent Systems«, *in Proceedings of Autonomous Agents '99 (Workshop on Deception, Fraud and Trust in Agent Societies).* May 1999, Seattle, pp. 149-161. [http://www.lb.cs.cmu.edu/~softagents/papers/aa99-addingSecurity.ps].

YU ET AL. 2000
Ting Yu, Xiaosong Ma, Marianne Winslett: »PRUNES: An Efficient and Complete Strategy for Automated Trust Negotiation over the Internet«, to appear in *ACM Conference on Computer and Communications Security, Athens,* November 2000. [http://drl.cs.uiuc.edu/pubs/ccs2000.ps].

YU / SINGH 2000
Bin Yu and Munindar P. Singh: »A Social Mechanism of Reputation Management in Electronic Communities«, in *Proceedings of the 4th International Workshop on Cooperative Information Agents (CIA).* Berlin: Springer 2000.

ZACHARIA 1999
Giorgos Zacharia: »Trust management through reputation mechanisms«, in C. Castelfranchi et al. (eds.): *Deception, Fraud and Trust in Agent Societies,* Seattle 1999, pp. 163-167.

ZACHARIA ET AL. 1999
Giorgos Zacharia, Alexandros Moukas and Pattie Maes: »Collaborative Reputation Mechanisms in Electronic Marketplaces«, in *Proceedings of the 32nd Hawaii International Conference on System Sciences,* Wailea Maui 1999.