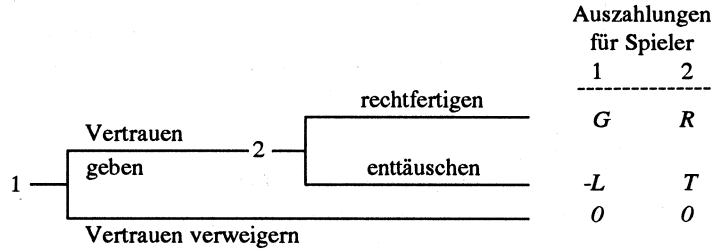
Werner Raub

Eine Notiz über die Stabilisierung von Vertrauen durch eine Mischung von wiederholten Interaktionen und glaubwürdigen Festlegungen*

Abstract: Various mechanisms are known that can stabilize trust relations. Examples are repeated interactions and credible commitments through warranties, deposits, and other kinds of 'hostages'. Usually, these mechanisms are studied in isolation from one another. An integrated analysis is widely neglected. In this note, the effects of a 'mix' of mechanisms are analyzed. A simple case is offered, where a combination of repeated interactions and credible commitments can stabilize trust, while neither of the mechanisms alone can do so.

Vertrauensbeziehungen sind ein zentrales Thema in Colemans *Grundlagen der Sozialtheorie* (z.B. Kapitel 5, 8, 28). Coleman diskutiert eine Reihe von Beispielen für Vertrauensbeziehungen und skizziert in intuitiver Weise verschiedene Mechanismen, die in solchen Beziehungen die Gewährung von Vertrauen durch einen Treugeber und die Rechtfertigung des Vertrauens durch einen Treuhänder stützen können. Ein nützliches spieltheoretisches Modell einfacher Vertrauensbeziehungen, wie sie v.a. in Kapitel 5 von Colemans Werk untersucht werden, ist das *Vertrauensspiel* (Trust Game) von Dasgupta (1988) und Kreps (1990).

Abbildung1: Das Vertrauensspiel



Annahmen über Auszahlungen: $G, L > 0$ und $T > R > 0$.

---

* Für nützliche Kommentare ist Chris Snijders zu danken. Eine englischsprachige Fassung entstand während eines von der Earhart-Foundation finanzierten Gastaufenthalts am Department of Sociology der University of Chicago.

Analyse & Kritik 14 (1992), S.187-194

Im Vertrauensspiel ist Spieler 1 der Treugeber. Er kann Vertrauen geben oder aber Vertrauen verweigern. Die Interaktion endet, wenn Vertrauen verweigert wird. Falls Vertrauen gegeben wird, ist Spieler 2, der Treuhänder, am Zug. Er kann Vertrauen rechtfertigen oder aber enttäuschen. Die Ordnung der Auszahlungen (kardinale Nutzenwerte) zeigt, daß der attraktivste Ausgang für den Treugeber gerechtfertigtes Vertrauen ist, gefolgt von verweigertem Vertrauen und enttäuschtem Vertrauen. Für den Treuhänder ist es am vorteilhaftesten, gegebenes Vertrauen zu enttäuschen, gefolgt von der Rechtfertigung von gegebenem Vertrauen und von verweigertem Vertrauen. Wir nehmen an, daß Spielbaum und Auszahlungen 'common knowledge' der Spieler sind, d.h. grob gesagt, daß Spielbaum und Auszahlungen beiden bekannt sind und daß dies auch beide wissen.[1]

Es ist offensichtlich, daß das Vertrauensspiel ein eindeutiges (teilspielperfektes) Gleichgewicht derart hat, daß Spieler 1 Vertrauen verweigert, während Spieler 2 Vertrauen enttäuschen würde: dies ist die einzige Strategienkombination, bei der jeder Spieler seine eigene Auszahlung maximiert, gegeben die Strategie des anderen Spielers.[2] Dieses Gleichgewicht ist ineffizient im Sinn des Pareto-Kriteriums: verglichen mit dem Ausgang 'verweigertes Vertrauen' stehen sich beide Spieler besser, wenn Vertrauen gegeben und gerechtfertigt wird.

Zwei bekannte *Mechanismen, die Vertrauensbeziehungen* zwischen rationalen Akteuren in dem Sinn *stabilisieren* können, daß Vertrauen gegeben und gerechtfertigt wird, sind wiederholte Interaktionen und glaubwürdige Festlegungen durch Garantien und andere Arten von Pfändern. Beide Mechanismen werden auch von Coleman (Kapitel 5) thematisiert, wenn er z.B. die langfristigen Folgen hervorhebt, die die Enttäuschung von Vertrauen für den Treuhänder nach sich ziehen kann, oder wenn er vertragliche Konstruktionen zur Absicherung von Vertrauensbeziehungen streift.

Wir betrachten zunächst den Fall *wiederholter Interaktionen* (siehe für eine technische Übersicht über die Theorie wiederholter Spiele Fudenberg/Maskin 1986 und für einflußreiche Anwendungen Taylor 1976/1987 und Axelrod 1987). Wir nehmen an, daß das Vertrauensspiel unbestimmt oft (technisch: unendlich oft) wiederholt wird in den Runden $t = 1, 2, \ldots$. Nach jeder Runde t werden beide Spieler informiert über das Verhalten des Partners in dieser Runde. Für

---

[1] Eine naheliegende Komplikation des Modells ergibt sich, wenn man annimmt, daß Spieler 1 nicht genau weiß, ob für Spieler 2 die Enttäuschung von Vertrauen attraktiver ist als die Rechtfertigung von Vertrauen. Diese Annahme führt zu einem Spiel mit unvollständiger (incomplete) Information, das bei Dasgupta 1988 besprochen wird. Eine spieltheoretische Fundierung von Colemans (1991, 126) Bedingung für die Vergabe von Vertrauen läßt sich in einfacher Weise aus einem solchen Spiel mit unvollständiger Information gewinnen.

[2] Die Erläuterung im Text bezieht sich auf ein einfaches (Nash) Gleichgewicht. Zum Begriff der Teilspielperfektheit, der Standard-Verschärfung des Gleichgewichtsbegriffs, vgl. Selten 1965.

jeden Spieler i ist seine Auszahlung U_i für das wiederholte Spiel die diskontierte Summe seiner Auszahlungen u_{it} in den einzelnen Runden, d.h.

$$U_i = \sum_{t=1}^{\infty} w^{t-1} u_{it},$$

wobei $0 < w < 1$ für den Diskontparameter w.

Aus dem 'Folk Theorem' für wiederholte Spiele folgt, daß es für ausreichend große Diskontparameter w teilspielperfekte Gleichgewichte im wiederholten Vertrauensspiel gibt, so daß in jeder Runde Vertrauen gegeben und gerechtfertigt wird. Diese Gleichgewichte können, grob gesprochen, aus bedingten Strategien bestehen, die vorschreiben, in jeder Runde Vertrauen zu geben (zu rechtfertigen), falls in allen vorherigen Runden Vertrauen gegeben und gerechtfertigt wurde, und andererseits nach dem ersten Fall von verweigertem oder enttäuschtem Vertrauen in allen zukünftigen Runden Vertrauen zu verweigern (zu enttäuschen). Der zentrale Effekt des Gebrauchs solcher bedingten Strategien ist, daß der Treuhänder den kurzfristigen Gewinn aus der Enttäuschung von Vertrauen im Vergleich zur Rechtfertigung von Vertrauen abwägen muß gegen die langfristigen Kosten der Enttäuschung von Vertrauen, die dadurch entstehen, daß der Treugeber in der Zukunft kein Vertrauen mehr gewährt. Wenn solche Gleichgewichte existieren, können wiederholte Interaktionen ausreichen, um Vertrauensbeziehungen zu stabilisieren. Wenn allerdings der Diskontparameter klein ist, genauer, wenn

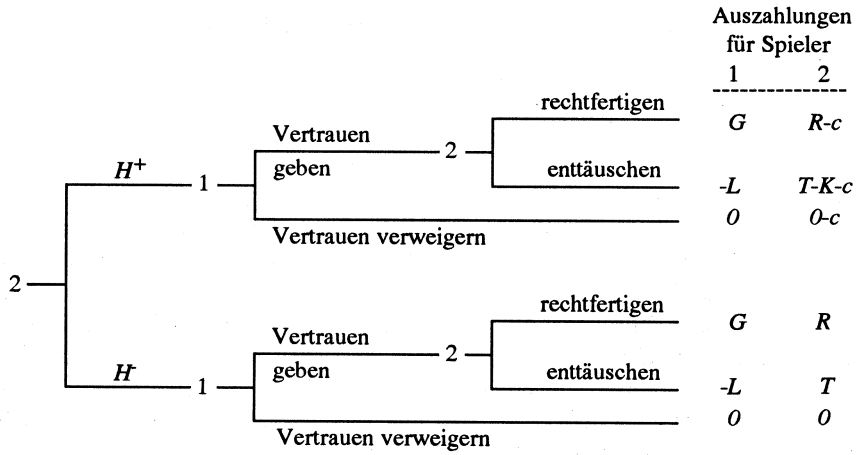
$$w < 1 - R/T, \tag{1}$$

dann gibt es kein Gleichgewicht für das wiederholte Vertrauensspiel, welches für jede Runde den Ausgang 'gerechtfertigtes Vertrauen' erzeugt.

Eine Alternative für die Stabilisierung von Vertrauensbeziehungen sind *glaubwürdige Festlegungen* durch Garantien und andere Arten von *Pfändern* (Weesie/Raub 1992). Abbildung 2 zeigt ein *Pfänderspiel*. Dabei handelt es sich um ein erweitertes Vertrauensspiel, bei dem der Treuhänder ein Pfand geben kann.

In diesem Spiel macht Spieler 2, der Treuhänder, den ersten Zug. Er kann ein Pfand ('hostage') geben, notiert als H^+, oder aber die Stellung eines Pfandes verweigern (H^-). Der Wert des Pfandes für Spieler 2 sei K. Wenn kein Pfand gegeben wird, spielen die beiden Spieler im Anschluß das ursprüngliche Vertrauensspiel. Wenn ein Pfand gegeben wird, spielen sie eine Variante des ursprünglichen Vertrauensspiels mit veränderten Auszahlungen. Genauer gesagt ist die Stellung eines Pfandes für Spieler 2 zunächst mit (Transaktions-) Kosten $c > 0$ verbunden: er muß z.B. einen Anwalt honorieren, der das Pfand in Verwahrung nimmt und er hat Opportunitätskosten, weil er das Pfand nicht für andere Zwecke

Abbildung 2: Das Pfänderspiel



verwenden kann. Wir nehmen an, daß diese Kosten unabhängig sind von den weiteren Zügen der beiden Spieler. Wenn Spieler 1 nach der Stellung eines Pfandes Vertrauen gibt und Spieler 2 dieses Vertrauen enttäuscht, dann verliert Spieler 2 sein Pfand, andernfalls erhält er es zurück.[3] Zu beachten ist, daß Spieler 1 vor seinem eigenen Zug darüber informiert ist, ob Spieler 2 ein Pfand gestellt hat oder nicht.

Wir wollen den Fall ausschließen, daß die Stellung eines Pfandes für Spieler 2 schon deshalb unattraktiv ist, weil die damit verbundenen Kosten zu hoch sind. Wir nehmen deshalb an, daß

$$R > c. \tag{2}$$

Man überzeugt sich leicht (Weesie/Raub 1992, Theorem 1), daß das Pfänderspiel genau dann ein teilspielperfektes Gleichgewicht derart hat, daß Spieler 2 ein Pfand stellt, Spieler 1 Vertrauen gibt und Spieler 2 Vertrauen rechtfertigt, wenn $K \geq T - R$. Unter dieser Bedingung wirkt die Stellung eines Pfandes als glaubwürdige Festlegung, die die Vertrauensbeziehung stabilisiert. Die Gleichgewichtsstrategien der Spieler machen dabei das Verhalten im Vertrauensspiel selbst abhängig von der vorangegangenen Stellung eines Pfandes durch Spieler 2. Demgegenüber hat das Pfänderspiel ein eindeutiges teilspielperfektes Gleichgewicht derart, daß kein Pfand gestellt und kein Vertrauen gegeben wird, wenn

$$T - R > K. \tag{3}$$

---

[3] Es macht keine Mühe, die Analyse auf komplexere 'Pfänderinstitutionen' auszudehnen (Weesie/Raub 1992), bei denen ein verfallenes Pfand von Spieler 2 z.B. an Spieler 1 gegeben wird.

Wir wollen nun zeigen, daß Vertrauensbeziehungen durch eine *Kombination* von wiederholten Interaktionen und glaubwürdigen Festlegungen durch Pfänder selbst dann stabilisiert werden können, wenn keiner der beiden Mechanismen allein für die Stabilisierung ausreicht. Dazu betrachten wir das unbestimmt oft wiederholte Pfänderspiel, wobei als Auszahlung für das wiederholte Spiel wiederum die diskontierte Summe der Auszahlungen in den einzelnen Runden angenommen wird. Wir nehmen außerdem erneut an, daß jeder Spieler nach jeder Runde über alle Züge des Partners in dieser Runde informiert ist. Sofern nun Ungleichung (1) gilt, gibt es kein Gleichgewicht im wiederholten Pfänderspiel, so daß Spieler 2 in keiner Runde ein Pfand stellt, während in jeder Runde Spieler 1 Vertrauen gibt und Spieler 2 Vertrauen rechtfertigt. Andererseits gilt unter den Ungleichungen (2) und (3) aber auch, daß es im wiederholten Pfänderspiel kein Gleichgewicht gibt, so daß in jeder Runde Spieler 2 ein Pfand stellt, Spieler 1 Vertrauen gibt und Spieler 2 Vertrauen rechtfertigt, wobei Spieler 1 sein Verhalten in einer gegebenen Runde ausschließlich abhängig macht von der Stellung eines Pfandes durch Spieler 2 *in dieser Runde*, nicht aber vom Verlauf des wiederholten Pfänderspiels *in früheren Runden*.

Man betrachte nun die folgende Strategie s_1 für Spieler 1:

- Gib Vertrauen in Runde 1, wenn ein Pfand gestellt wurde.
  Verweigere Vertrauen in Runde 1, wenn kein Pfand gestellt wurde.
- Gib Vertrauen in Runde $t = 2, 3, \ldots$ wenn die folgenden drei Bedingungen erfüllt sind:
  (i) In allen Runden $1, \ldots, t$ wurde ein Pfand gestellt.
  (ii) In allen vorangegangenen Runden $1, \ldots, t - 1$ wurde Vertrauen gegeben.
  (iii) In allen vorangegangenen Runden $1, \ldots, t - 1$ wurde Vertrauen gerechtfertigt.
  Verweigere Vertrauen in allen anderen Fällen.

Für Spieler 2 betrachte man die folgende Strategie s_2:

- Stelle ein Pfand in Runde 1.
- Stelle ein Pfand in Runde $t = 2, 3, \ldots$, wenn die folgenden drei Bedingungen erfüllt sind:
  (i) In allen vorangegangenen Runden $1, \ldots, t - 1$ wurde ein Pfand gestellt.
  (ii) In allen vorangegangenen Runden $1, \ldots, t - 1$ wurde Vertrauen gegeben.
  (iii) In allen vorangegangenen Runden $1, \ldots, t - 1$ wurde Vertrauen honoriert.
  Stelle andernfalls kein Pfand.
- Rechtfertige Vertrauen in Runde 1, wenn in Runde 1 ein Pfand gestellt und Vertrauen gegeben wurde. Enttäusche Vertrauen andernfalls.
- Rechtfertige Vertrauen in Runde $t = 2, 3, \ldots$, wenn die folgenden drei Bedingungen erfüllt sind:

(i) In allen Runden 1, ..., t wurde ein Pfand gestellt.
(ii) In allen Runden 1, ..., t wurde Vertrauen gegeben.
(iii) In allen vorherigen Runden 1, ..., t - 1 wurde Vertrauen gerechtfertigt.
Enttäusche Vertrauen in allen anderen Fällen.

Offensichtlich wird im wiederholten Pfänderspiel in jeder Runde ein Pfand gestellt, Vertrauen gegeben und Vertrauen gerechtfertigt, wenn s_1 und s_2 gespielt werden.

Theorem: Es sei angenommen, daß die Ungleichungen (1) - (3) gelten. Dann ist $s = (s_1,s_2)$ genau dann ein teilspielperfektes Gleichgewicht des wiederholten Pfänderspiels, wenn die beiden folgenden Bedingungen erfüllt sind:

$$KR / (T - R) > c \tag{4}$$

und

$$w \geq (T - K - R) / (T - K - c). \tag{5}$$

[ ]

Beweis: Man beachte zunächst, daß Ungleichung (4) sicherstellt, daß (1) und (5) für geeignete Parameter des Spiels simultan erfüllt werden können. Es ist außerdem klar, daß s_1 die Auszahlung von Spieler 1 gegen s_2 maximiert, weil Spieler 1 in jeder Runde seine maximale Auszahlung G erhält. Es bleibt zu prüfen, wann s_2 die Auszahlung von Spieler 2 gegen s_1 maximiert. Die mit s verbundene Auszahlung für Spieler 2 ist $u_2(s) = (R - c) / (1 - w)$. Wenn Spieler 2 überhaupt seine Auszahlung durch Abweichung von s_2 verbessern kann, dann ist bereits eine Abweichung in Runde 1 attraktiv. Wenn er in der ersten Periode kein Pfand stellt, dann ist seine Auszahlung für das wiederholte Spiel $u^*_2 = 0 < u_2(s)$. Wenn Spieler 2 in Runde 1 ein Pfand stellt und anschließend Vertrauen enttäuscht, erhält er eine Auszahlung $u^{**}_2 = T - K - c$ für das wiederholte Spiel. Nun gilt $u_2(s) \geq u^{**}_2$ genau dann, wenn Ungleichung (5) erfüllt ist. Wenn also (5) gilt, maximiert s_2 die Auszahlung von Spieler 2 gegen s_1 und s ist Gleichgewicht. Die Teilspielperfektheit des Gleichgewichts folgt daraus, daß unter (3) im nicht wiederholten Pfänderspiel ein eindeutiges teilspielperfektes Gleichgewicht existiert, so daß kein Pfand gestellt, Vertrauen verweigert und Vertrauen enttäuscht wird. [ ]

Beispiel: Als numerisches Beispiel betrachte man ein wiederholtes Pfänderspiel mit $T = 5$, $R = 3$, $K = c = 1$ und $w = 11/30$. Für dieses Spiel ist s ein teilspielperfektes Gleichgewicht. Es gibt demgegenüber kein teilspielperfektes Gleichgewicht derart, daß Spieler 1 seine Vertrauensgewährung *nur* abhängig macht von der Rechtfertigung von Vertrauen durch Spieler 2 in allen früheren Runden. Ebenso gibt es kein teilspielperfektes Gleichgewicht derart, daß Spieler 1 die Vertrauensgewährung in Runde t *nur* abhängig macht von der Stellung eines Pfandes durch Spieler 2 in dieser Runde t.

Die hier vorgelegte Skizze ist ersichtlich nicht mehr als ein bescheidenes erstes Beispiel für die Durchführung eines umfangreicheren *Forschungsprogramms* (siehe zu diesem Programm Raub/Weesie 1992) zur Kooperation rationaler und eigeninteressierter Akteure in mit Anreizproblemen behafteten 'problematischen Situationen' (Voss 1985), bei denen individuell rationales Verhalten zu kollektiv irrationalen (Pareto-ineffizienten) Ergebnissen führen kann. Eine Verallgemeinerung der Analyse müßte einerseits neben dem Vertrauensspiel andere Typen solcher problematischen Situationen abdecken. Ein zweiter Verallgemeinerungsschritt würde darin bestehen, weitere Mechanismen in die Analyse einzubeziehen, die in problematischen Situationen individuell und kollektiv rationales Verhalten zur Deckung bringen können. Zu denken ist etwa an Reputationseffekte, die sich ergeben, wenn z.B. die Enttäuschung gewährten Vertrauens durch den Treuhänder auch anderen Partnern des Treuhänders bekannt wird und deren zukünftiges Verhalten dem Treuhänder gegenüber beeinflußt (siehe Raub/Weesie 1990). Zu denken ist auch an die Effekte von Exit-Optionen, die es ermöglichen, die Enttäuschung gewährten Vertrauens nicht durch direkte Sanktionen, sondern durch die Aufkündigung zukünftiger Interaktionen abzuschrecken (siehe z.B. Schuessler 1990; Weesie 1992). Ein entscheidender dritter Typ von Verallgemeinerungen der Analyse würde schließlich auf eine Endogenisierung der Mechanismen abstellen. Damit ist eine Untersuchung der Bedingungen gemeint, unter denen eigeninteressiert handelnde Akteure von Möglichkeiten Gebrauch machen, sich selbst ex ante solche Mechanismen zu schaffen, die ex post Anreizprobleme und ineffiziente Ausgänge von Interaktionen verhindern. Zu denken ist etwa an Szenarien, in denen Akteure zwischen verschiedenen Arten von Pfändern wählen können oder in denen es, jedenfalls in gewissem Umfang, den Akteuren selbst überlassen ist, etwa eine komplexe Transaktion in einzelnen Schritten abzuwickeln und so für wiederholte Interaktionen zu sorgen.

Bibliographie

Axelrod, Robert (1987), *Die Evolution der Kooperation*, München

Coleman, James S. (1991-93), *Grundlagen der Sozialtheorie. 3 Bde.*, München

Dasgupta, Partha (1988), Trust as a Commodity, in: Diego Gambetta (ed.), *Trust - Making and Breaking Cooperative Relations*, Oxford, 49-72

Fudenberg, Drew/Eric Maskin (1986), The Folk Theorem in Repeated Games with Discounting or with Incomplete Information, in: *Econometrica 54*, 533-554

Kreps, David M. (1990) Corporate Culture and Economic Theory, in: James E. Alt/ Kenneth A. Shepsle (eds.), *Perspectives on Positive Political Economy*, Cambridge, 90-143

Raub, Werner/Jeroen Weesie (1990), Reputation and Efficiency in Social Interactions: An Example of Network Effects, in: *American Journal of Sociology 96*, 626-654

- / - (1992), *The Management of Matches*, mimeo, Utrecht University

Schuessler, Rudolf (1990), *Kooperation unter Egoisten: vier Dilemmata*, München

Selten, Reinhard (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit, in: *Zeitschrift für die gesamte Staatswissenschaft 121*, 301-324 und 667-689

Taylor, Michael (1976/1987) *The Possibility of Cooperation*, Cambridge (rev. Aufl. von *Anarchy and Cooperation*, London)

Voss, Thomas (1985), *Rationale Akteure und soziale Institutionen. Beitrag zu einer endogenen Theorie des sozialen Tauschs*, München

Weesie, Jeroen (1992), *Disciplining via Exit and Voice*, mimeo, Utrecht University

- /Werner Raub (1992), Private Ordering. *A Comparative Institutional Analysis of Hostage Games*, paper prepared for the Annual Meeting of the American Sociological Association, Pittsburgh/PA