.

Trust as Encapsulated Interest

Russell Hardin (2002)

1 Hardins Vertrauensdefinition

> To say that I trust you in some context simply means that I think you will be *trustworthy* toward me in that context. (S.1)

Dabei mekrt Hardin an, dass Vertrauen immer *relational* ist. Damit meint er, dass Vertrauen nur innherlab interpersoneller Beziehungen möglich ist.[1] Wann aber gilt jemand als *vertrauenswürdig*?

2 Encapsulated Interest

Der wichtigste und häufigste (nicht aber der einzige) Grund für eine Person A, anzunehmen, dass eine Person B vertrauenswürdig ist, ist Encapsulated Interest:

> I trust you because I think it is in your interest to take my interests in the relevant matter seriously in the following sense: You value the continuation of our relationship, and you therefore have your own interests in taking my interests into account. That is, you encapsulate my interests in your own interests.

Anmerkungen:

- Vertrauenswürdig erscheinen vs. ververtrauenswürdig sein: Damit eine Person X einer Person Y vertraut, reicht es aus, dass A B für vertrauenswürdig *hält*. In anderen Worten: As Überzeugung, dass B vertrauenswürdig ist, ist hinreiched dafür, dass A B vertraut. Das bedeutet, B muss nicht tatsächlich vertrauenswürdig *sein*, damit A B vertraut. B ist nämlich nur dann vertrauenswürdig, wenn die Proposition in As Überzeugung wahr ist.
- Vertrauenswürdigkeit vs. Handeln: Dass B im Interesse von A *handelt*, ist weder notwendig dafür, dass B vertrauenswürdig ist, noch dafür, dass B für vertrauenswürdig gehalten wird: Neben dem encapsulated interest von A (als Interesse von B) kann B noch weitere Interessen haben, die das encapsulated interest übertrumpfen.
- Gleiches Interesse zweier Personen: Dass zwei Personen (zufällig) das gleiche Interesse haben, erfüllt *nicht* die Kriterien für encapsulated interest. Die Person, der vertraut wird, muss die Beziehung mit der Person, die ihr vertraut, so wertschätzen, dass sie sie weiterführen möchte und aus diesem Grund die Interessen der vertrauenden Person in ihre Überlegungen miteinbezieht.

3 Elemente von Vertrauen als Ecapsulated Interest

Hauptelemente (vgl. S.7ff.):

- Dreistrelligkeit von Vertrauen: Person **A** vertraut Person **B** Φ zu tun.
- Vertrauen als Cognitive Notion: Ebenso wie *Wissen* und *Überzeugung* ist Vertrauen eine *cognitive notion*. Sie alle basieren auf die ein oder andere Weise auf einem Verständnis davon, was wahr ist. Dementsprechend kann sich niemand entscheiden, zu vertrauen, ebenso, wie niemand über die Wahrheit entscheiden kann (Wahrheit kann nur entdeckt werden bzw. man kann davon überzeugt sein, dass etwas wahr/falsch ist). Eine Person kann sich also nicht dazu entscheiden, einer anderen Person zu vertrauen. Sehr wohl aber kann sie sich dazu entscheiden, vertrauenswürdig zu sein. (vgl. S. 7)

[1] Er unterscheidet hier zwischen direkter Interaktion und Vermittlern/Reputationseffekten.

- Risiko: Handlungen, die auf Vertrauen basieren, invovieren ein gewisses Risiko.[2] In anderen Worten: Von Vertrauen zu sprechen, ist nur im Angesicht einer Unsicherheit sinnvoll.

Weitere Elemente:

- Erwartungen über Verhalten: Vertrauen beinhaltet Erwartungen über das Verhalten einer anderen Person.
- Kompetenz: Person A kann Person B nur vertrauen, wenn A B in Φ für kompetent hält. (vgl. oben: Unterschied kompetenz *erscheinen* vs. kompetent *sein*)
- Jenseits von interpersonellem Vertrauen: In unseren Beziehungen mit Fachleuten in institutionalisierten Kontexten (z.B. mit Doktor:innen, Anwält:innen, etc.) denken und handeln wir eher auf Basis von Einschätzungen Dritter und bloßen Erwartungen. Diese ersetzen zu einem großen Teil das Vertrauensverhältnis, das wir auf einer interpersonellen Ebene normalerweise haben.[3]

4 Interpersonelles Vertrauen als Iteration des Gefangenendilemmas

Einführung in das Gefangenendilemma: SEP Prisoner's Dilemma

Prototypische Fälle von gegenseitigem Vertrauen (*mutual trust*) (**(!)** nicht jedoch alle Fälle von Vertrauen **(!)**) lassen sich im Grunde genommen in jedem Durchlauf (also in jeder vertrauensbasierten Interaktion) wie ein Gefangenengendilemma beschreiben:

Im klassischen Gefangenendilemma geben Zahlen an, wie viele Jahre zwei eines Verbrechen beschuldigte Personen jeweils ins Gefängnis müssen, abhängig davon, ob sie die Tat gestehen/schweigen/die jeweils andere Person belasten. In seiner abstrakten Form (siehe unten) ordnen die Zahlen jedoch lediglich die verschiedenen Ergebnisse und werden deshalb nicht selten durch Buchstaben ersetzt. Sie geben ein relatives Verhältnis bzw. eine Gewichtung der Outcomes an:

$\downarrow A/B \rightarrow$	C (cooperate)	D (defect)
C	R_A, R_B	S_A, T_B
D	T_A, S_B	P_A, P_B

Dabei gilt: T >[4] R > P >S
temptation > reward > punishment > sucker's payoff

Bei nur einem einzigen Durchlauf haben beide Personen aus ihrer eigenen Perspektive den größten Vorteil, wenn sie egoistisch sind und der anderen Person nicht vertrauen (**D** für *defect*). Denn dann genießen sie den Vorteil, den sie davon haben, dass ihnen die andere Person vertraut , ohne selbst vertrauen zu müssen bzw. eine daraus resultierende Handlung (Gegenleistung) durchführen zu müssen (**T** für *temptation*). Oder sie gehen leer aus, wenn die andere Person sich entscheidet, ebenfalls nicht zu kooperieren (**P** für *punishment*). Sie können also etwas gewinnen, ohne das Risiko zu haben, etwas zu verlieren. Sollte sich eine der Personen entscheiden, zu vertrauen (**C** für *cooperate*), kann das Vertrauen von der anderen Person erwidert werden. Beide haben in diesem Fall "Kosten" in Form der Handlung/der Verwundbarkeit/des Risikos, das mit dem Vertrauen einhergeht, gewinnen allerdings ebendies von der anderem Person, die ihr Vertrauen erwidert (**R** für *reward*). Sollte sich die andere Person jedoch entscheiden, nicht zu vertrauen, bleiben die "Kosten" bestehen, nicht aber die Vorteile, die sich aus dem Vertrauen der anderen Person ergeben hätten (**S** für *sucker's payoff*). Entscheidet man sich also dafür, zu vertrauen, gewinnt man entweder etwas, das mit "Kosten" verbunden ist, oder man verliert etwas. Im direkten Vergleich der beiden Handlungsoptionen

[2] Der Akt des Vertrauens selbst invovliert dabei nicht das Risiko, sondern erst das Handeln auf der Basis des Vertrauens.
[3] vgl. Hardins Konzept von Quasi-Vertrauen in Hardin, Russell (2002): "Trust and Government"
[4] gelesen als: ist für das jeweilige Individuum besser als

scheint es zunächst attraktiver, etwas gewinnen zu können, ohne das Risiko, etwas zu verlieren, als etwas zu gewinnen, das gleichzeitig Kosten verursacht oder etwas zu verlieren. Damit haben beide Personen einen Anreiz, sich nicht zu vertrauen.
Wenn sich nun allerdings beide Personen nicht vertrauen, landen sie bei dem Ergebnis, das, absolut gesehen, am schlechtesten ist: Sie bekommen beide gar nichts . Vertrauen sich die Personen gegenseitig , profitieren sie jeweils vom Vertrauen der anderen Person , müssen aber im Austausch für das erhaltene Vertrauen und die damit möglicherweise einhergehende Handlung, das Vertrauen und die Handlung erwidern. Bei nur einem Durchlauf ist für jedes Individuum der Vorteil durch **T** jeweils größer als durch **R**, da es in beiden Fällen vom Vertrauen profitiert, aber nur im zweiten Fall eine Gegenleistung erbringen muss. Bei einer Iteration allerdings, kann sich **R** im Vergleich zu **T** als besser herausstellen, da in der Regel einmal missbrauchtes Vertrauen dazu führt, dass die vertrauende Person in Zukunft (in diesem Bereich) der missbrauchenden Person nicht wieder vertraut. Der absolute Vorteil der missbrauchenden Person beschränkt sich somit auf **T**. Wenn sich aber beide Personen wieder und wieder vertrauen (die Iteration ist also entweder unendlich (*infinite iteration*) oder wird auf unbezstimmte Zeit fortgesetzt (*indefinite iteration*), können beide in jedem Schritt der Iteration den Vorteil **R** bekommen, der bei genügend Wiederholungen in der Summe größer sein wird als **T**. Ist die Iteration jedoch endlich und die beteiligten Parteien wissen um die Endlichkeit, ergibt sich das Problem der *backwards induction* (siehe SEP Artikel). In langfristigen Beziehungen haben also beide Parteien einen Anreiz, zu vertrauen.

5 Fragen

1. In welchen Fällen würden wir von Vertrauen sprechen, auf die sich Hardins Encapuslated-Interest-Modell nicht anwenden lässt? Was sagt Hardin dazu?

2. Was meint Hardin, wenn er sagt, Vertrauen sei eine *cognitive notion*? (vgl. S. 7)

3. Hardin merkt an, dass Spieltheoretiker:innen dafür argumentieren, dass eine Iteration im klassischen Gefangenendilemma keine Anreize schaffen kann, wie in 4. beschrieben. Wie funktioniert ihr Argument der Rückwärtsinduktion laut Hardin? (vgl. S.18, 20)

4. Ist Hardins Ansatz, gängige Fälle gegenseitigen Vertrauens spieltheoretisch als Iteration des Gefangenendilemmas aufzufassen, plausibel? Welche Aspekte von Vertrauen fängt er damit *nicht* ein?

Referenz

Russell, Hardin (2002): “Trust”, *Trust and Trustworthiness*, New York: Russell Sage Foundation, pp. 1-27.