Das Gefangenendilemma[1]

Das Gefangenendilemma ist der „Beißknochen“ der Spieltheorie – man kann endlos auf ihm herumkauen. Tausende von Mathematikern, Psychologen, Politologen, Philosophen und Volkswirtschaftlern haben sich mit ihm beschäftigt und versucht, das Dilemma zu lösen, und trotzdem ist es noch immer so geheimnisvoll und verblüffend wie 1950, als Merrill Flood und Melvin Drescher es vorstellten. Seinen Namen erhielt es von Albert W. Tucker, der 1951 die erste Arbeit darüber schrieb. Tucker kleidete das Problem in Form eines kurzen Krimis – jeder, der seither darüber geschrieben hat, hat es in anderen Farben ausgemalt. Hier ist eine Fassung:

Die Polizei fängt ein langgesuchtes Duo, dem ein schweres Verbrechen angelastet wird. Die Polizei hat keine direkten Beweise für die Schuld der beiden Männer und kann ihnen lediglich Fahren mit überhöhter Geschwindigkeit nachweisen. Der Untersuchungsrichter möchte den Fall abschließen und unterbreitet dazu jedem der Gefangenen, die er in getrennte Zellen legen ließ, den folgenden Vorschlag:

„Wenn Sie das Verbrechen gestehen und uns dadurch helfen, den Fall zu klären, lasse ich Sie frei, und wir vergessen die Sache mit dem zu schnellen Fahren. Ihren Komplizen lochen wir dann für zehn Jahre ein, und die Sache ist für immer erledigt. Das Angebot gilt jedoch nur, falls Ihr Komplize das Verbrechen nicht gesteht und uns also bei der Aufklärung nicht hilft. Wenn er ebenfalls gesteht, ist Ihr Geständnis nicht viel wert, weil wir dann ohnehin Bescheid wissen. In diesem Fall werden Sie beide zu je fünf Jahren Gefängnis verurteilt. Wenn keiner von Ihnen gesteht, ist das Ihre Sache, aber in dem Fall wird diese fürchterliche Raserei streng geahndet werden, und Sie kommen beide für ein Jahr hinter Schloß und Riegel. Noch etwas: Ihrem Komplizen habe ich das gleiche gesagt wie Ihnen. Ich erwarte Ihre Antwort morgen früh um zehn Uhr – um elf Uhr können Sie frei sein.“

Wir fassen die Situation in einer Matrix zusammen:

		Der andere Komplize: gesteht	Der andere Komplize: gesteht nicht
Der erste Komplize	gesteht	–5, –5	0, –10
	gesteht nicht	–10, 0	–1, –1

Die erste, unterstrichene Zahl in jedem Feld zeigt das „Ergebnis“ für den ersten Komplizen, die zweite das für den anderen. Da es schlimmer ist, zehn Jahre abzusitzen als fünf, müssen die Haftjahre „negativ“ gezählt werden. In der gegebenen Situation ist Null das bestmögliche Ergebnis.

Zwei logische Lösungen

Die beiden Komplizen hegen keinerlei Gefühle füreinander, ihre Zusammenarbeit war rein zufällig. Beide haben lediglich das Ziel, so billig wie möglich davonzukommen. Was ist für sie die logische Lösung: gestehen oder nicht?

Versetzen wir uns in die Rolle eines der Komplizen, und versuchen wir, aus seiner Sicht logisch zu denken: Wenn mein Partner singt, kann ich entweder auch singen und muß fünf Jahre brummen, oder ich gestehe nicht und muß zehn Jahre absitzen. Wenn mein Komplize gesteht, gestehe ich also besser auch.

Wenn mein Komplize nicht gesteht, gibt es wieder zwei Möglichkeiten: Ich gestehe und bin morgen frei, oder ich gestehe nicht und sitze ein Jahr ab. Wenn mein Komplize nicht gesteht, gestehe ich besser.

Ich komme also unabhängig davon, was mein Komplize macht, besser weg, wenn ich gestehe. Ich habe keine andere Wahl, weil mein Komplize nur zwei Möglichkeiten hat. Also befiehlt mir die Logik zu gestehen.

Die Logik diktiert dem anderen Gefangenen das gleiche. Weil nun beide logisch denken, gestehen sie beide – und jeder wird zu einer Haftstrafe von fünf Jahren verdonnert, obwohl sie mit nur einem Jahr davongekommen wären, wenn jeder dichtgehalten hätte. Das ist das *Gefangenendilemma*. Die Frage ist, ob diese Logik irgendwo fehlerhaft ist oder ob sie dieses für beide unangenehme Ergebnis erzwingt. In anderen Worten: Schließt die Logik rationale Zusammenarbeit der beiden Gefangenen aus?

[1] Mérö, László (2000): Die Logik der Unvernunft. 3. Auflage. Hamburg: Rowohlt, S. 47 ff.