HU Berlin WS 06/07
Seminar „Entscheidungstheorie. Einführung und philosophische Anwendungen“ von Thomas Schmidt
Essay von Johannes Stein

Was setzen das Gifträtsel und das Newcombsche Problem voraus?

Übersicht

Nach einer kurzen Behandlung des Newcombschen Problems werden wir uns ausführlicher dem Gifträtsel widmen,[1] wobei wir immer wieder auf Newcomb rekurrieren, um etwa Ähnlichkeiten und Unterschiede herauszustellen. Insbesondere soll untersucht werden, welchen Einfluss geistesphilosophische und metaphysische Überzeugungen auf den Ausgang der beiden Gedankenexperimente ausüben.

*

–„Wenn ich durch rationales Nachdenken über Newcombs Problem zu dem Ergebnis komme, nur eine Kiste zu nehmen, könnte ich genausogut beide nehmen, denn dann hätte das übernatürliche Wesen ja höchstwahrscheinlich vorhergesagt, dass ich nur eine nehmen würde. Aber Moment – bin ich nicht gerade durch rationales Nachdenken zu dem Ergebnis gekommen, beide Kisten zu nehmen? Und sollte das Wesen meine Überlegungen bis jetzt vorhergesagt haben, dann dürfte ihm auch der letzte Schritt keine Probleme bereitet haben.“ Und so entscheide ich mich, diesen Schritt zu verwerfen, und zu meinem vorherigen Entschluss zurückzukehren. Ich werde also nur eine Kiste nehmen. Obwohl ich dann ja beide Kisten nehmen könnte, *usw*. Wie man sieht, bekommt man große Schwierigkeiten, da man sich gedanklich nicht außerhalb des eigenen Denkens begeben kann. Sagt man sich etwa: „Das hat das Wesen vorausgesehen, was ergibt sich daraus für mich?“, so hat das Wesen diesen Gedanken höchstwahrscheinlich ebenfalls vorausgesehen. Es erweist sich sozusagen als unmöglich, die eigenen Gedanken durch angestrengtes Nachdenken einzuholen.

Eine solche Strategie scheint geringe Aussichten auf Erfolg zu haben. Werden wir also grundsätzlicher und fragen uns, ob die benötigte Vorhersagegenauigkeit überhaupt logisch möglich ist. Könnten wir dies irgendwie ausschließen, wären wir wieder Herr der Lage. Und tatsächlich: Das übernatürliche Wesen benötigt für seine Vorhersagen entweder eine sehr weitgehend determinierte Welt, in der unsere zukünftigen Entscheidungen bereits in unseren Gehirnen vorprogrammiert sind, oder aber die Möglichkeit, nach Belieben durch die Zeit zu reisen, und zu verschiedenen Zeitpunkten die Welt zu verändern (etwa das Geld aus der blauen Kiste zu nehmen, bzw. gar nicht erst hineingelegt zu haben, nachdem es beobachtet hat, dass wir beide Kisten genommen haben).

1 Kurzfassungen und Quellenverweise zu Newcombs Problem und dem Gifträtsel sind am Ende dieses Texts zu finden. Ihre Kenntnis wird im weiteren Verlauf vorausgesetzt.

Aus solcherart rückwärts gerichteter Kausalität ergeben sich jedoch logische Widersprüche, wie unzählige paradoxe Situationen im Zusammenhang mit Zeitreisen deutlich machen. Doch davon abgesehen können wir in einem weiteren Schritt sogar zugeben, dass all dies irgendwie funktionieren möge, doch könnten wir dann sowohl im Determinismusfall, als auch angesichts der Rückwärtskausalität den Begriff 'Entscheidung' nicht mehr sinnvoll verwenden.

**

Die Million im Gifträtsel zu bekommen scheint aus verschiedenen Gründen schwierig. Beginnen wir mit einer semantischen Einschränkung: Im üblichen Sprachgebrauch bedeutet eine Absicht (a) zu haben immer auch, die Meta-Absicht zu haben, (a) weiter bis zur Durchführung der durch (a) beabsichtigten Handlung beizubehalten. Es wäre also inkonsistent, wenn wir die Absicht zu hätten, das Gift zu trinken, und gleichzeitig nicht daran glaubten, dass wir es trinken würden, falls uns dies möglich wäre. Davon ausgehend zeichnen sich drei mögliche Handlungsverläufe ab, zwei davon irrational, einer inkonsistent:

1. Wenn man nicht glaubt, dass man das Gift trinken wird, kann man auch nicht beabsichtigen, es zu trinken. Wenn aber kein Grund besteht, es zu trinken, wäre das Trinken irrational. Also glaubt man nicht, dass man es trinken wird, und kann folglich auch nicht die Absicht haben, es zu trinken. Dies aber ist irrational, da man sich so eine Million Euro entgehen lässt.
2. Man nimmt sich fest vor, das Gift zu trinken, und muss sich deshalb am nächsten Tag durch das Trinken Schaden zufügen, obwohl es dann keinerlei Grund dazu gibt: Das Geld liegt auf meinem Tisch (oder nicht), und daran wird sich nichts ändern, egal ob ich das Gift trinke oder nicht. Also ist das Trinken irrational.
3. Zunächst beabsichtigt man ernsthaft, das Glas auszutrinken, am nächsten Tag nicht mehr. Diese Möglichkeit wahrt die Rationalität über die gesamte Zeit, opfert aber die Konsistenz.

Anscheinend können wir uns demnach aussuchen, entweder rational oder konsistent zu sein, nicht jedoch beides zugleich. Allerdings stellt sich die Frage, ob jede dieser Möglichkeiten überhaupt logisch widerspruchsfrei ist. Schließlich war das übernatürliche Wesen aus Newcombs Problem zunächst ebenfalls denkbar, bei genauerem Hinsehen jedoch logisch widersprüchlich. Stoßen wir vielleicht beim Gifträtsel auf ähnliche Schwierigkeiten? Die Durchführung der ersten Möglichkeit bedarf offensichtlich keiner großen Anstrengung, allerdings kommt sie einer Resignation vor der gestellten Aufgabe gleich. Betrachten wir deshalb die verbleibenden Optionen. Die zweite Möglichkeit geht bis an die Grenzen der Willenskraft eines Fanatikers, und sicherlich über die

durchschnittlichen hinaus, kann also ebenfalls verworfen werden.[1]

Bleibt die dritte Möglichkeit. Sie mag uns zunächst gangbar erscheinen, weil viele Menschen oft auf ähnliche Weise inkonsistent sind. Dies trifft allerdings nur unter der Prämisse zu, dass sich zwischenzeitlich ihr Wissenshorizont verändert hat. Die geforderte Inkonsistenz bei gleichbleibendem Wissensstand ist dagegen gleichsam schwieriger zu bewerkstelligen, als ein Gefühl der Spannung während einer Partie Go gegen sich selbst. Die Möglichkeit der diachronen Inkonsistenz bei gleichbleibendem Wissenshorizont erscheint mir als kaum weniger absurd, als eine synchrone Inkonsistenz der Art „Ich glaube, dass es (hier jetzt) schneit und nicht schneit". Eine derart konsequente Selbsttäuschung scheint für einen geistig gesunden Menschen äußerst schwierig, wenn nicht unmöglich zu sein. Aber könnte man nicht stattdessen das Messgerät täuschen?

***

Die übliche Interpretation des Gifträtsels widmet sich ausschließlich den Optionen des menschlichen Akteurs, wie wir es bisher ebenfalls getan haben. Mir scheint, dass uns ein anderer Fokus vielleicht ergänzenden Aufschluss bringen kann: Was ist das eigentlich für eine Maschine, die mich da analysiert? Kann es sie (aus logischer Sicht) überhaupt geben?

Unser Startpunkt für diesen Teil der Erkundung sei folgende Überlegung: Was wäre, wenn ich am entscheidenden Abend die feste Absicht hätte, das Gift am nächsten Tag zu trinken, das Messgerät aber aus irgendeinem Grund anzeigte, ich hätte diese Überzeugung nicht? Wie würde ich reagieren? Das Ausschlaggebende an dieser Stelle ist, was genau meine Absichten sind, welche das Messgerät analysiert. Dazu gibt es zwei grundlegend verschiedene Überzeugungen:

> Meine Absichten zum Zeitpunkt *t* sind eine (unechte) Teilmenge meiner bewussten Gedanken in *t*. (a)
>
> Meine Absichten sind wie (a), und zusätzlich eine (unechte) Teilmenge meiner Dispositionen, die sich nur als bewusste Absichten manifestieren, falls ich in eine geeignete Situation gerate. (b)

Falls (a), so könnte es mir durch Willensanstrengung gelingen, das Messgerät zu täuschen und die Million zu bekommen, ohne das Gift zu trinken. Dazu müsste ich während der Messung nur mantraartig denken: „Ich werde morgen das Gift trinken, ich werde morgen das Gift trinken, ich werde ...". Wäre hingegen (b) der Fall, zählten also auch jene Gehirnzustände, die nichts mit meinem augenblicklichen aktiven Denken zu tun haben, dann wäre meine Lage ziemlich

1 Da wir laut Vorgaben des Gedankenexperiments keine zusätzlichen Abmachungen treffen dürfen, können wir auch nicht etwa eine Zusatzwette mit hohem Einsatz darauf abschließen, dass wir das Gift trinken. Solches würde unser Problem beseitigen, da wir dann tatsächlich einen Grund zum Trinken hätten.

aussichtslos. In diesem Fall könnte ich während der Messung auch an Weihnachtsmänner denken oder sogar schlafen, ohne das Ergebnis zu beeinflussen. Wäre eine solche Messung möglich, so würde allerdings auch die Denkbarkeit des höheren Wesens von Newcomb, und mit ihm der Determinismus, in unangenehme Wirklichkeitsnähe rücken. Ich könnte dann nicht mehr mit Recht behaupten, das Gerät habe eine Störung und ich hätte eigentlich die 'richtige' Absicht gehabt – denn solch ein Gerät würde meine Absichten besser kennen als ich selbst.

Ich habe den Eindruck, dass die dritte Option aus dem vorhergehenden Abschnitt nur dann interessant bleibt, wenn das Gemessene weder (a) noch (b) entspricht. Wäre es (a), könnte man die Messung zu leicht manipulieren, unter Annahme (b) wäre eine Täuschung nicht denkbar. Mit (b) kämen außerdem weitere Unwägbarkeiten hinzu, wie folgender Fall illustriert: Nach Voraussetzung des Gedankenexperiments vertraue ich dem Messgerät, doch befreit mich dieses Vertrauen von einem Teil meiner Überzeugungen? Nehmen wir dazu an, das oben beschriebene Verhalten des Messgeräts träte auf (also keine gemessene trotz meiner gefühlten Überzeugung), und ich würde das intuitiv (aber wie beschrieben unberechtigtermaßen) als Fehlfunktion interpretieren. Wäre ich dann noch disponiert, das Gift zu trinken? Falls ich über diesen Fall noch nicht nachgedacht hatte, jedoch unter diesen Umständen nicht mehr bereit wäre, das Gift zu trinken, wäre dies dann nicht ein unzulässiges Abweichen von meiner Absicht, und damit für die Maschine Grund genug für ein negatives Testergebnis? Derlei Inkonsistenzen ließen sich in beliebiger Menge konstruieren. Sie legen nahe, dass Annahme (b) nicht unbedingt einen weitgehenden Determinismus zur Folge hat, sondern vielmehr das Verhalten der Maschine auch als eine Art rückwärtsgerichteter Kausalität interpretiert werden kann, wie sie uns schon bei der Untersuchung des Newcombschen Problems begegnet ist.

****

Man kann feststellen, dass die beiden untersuchten Gedankenexperimente nur scheinbar Beispiele für Probleme der Entscheidungstheorie sind. Vielmehr hängt ihr Ausgang von grundlegenden metaphysischen und geistesphilosophischen Überzeugungen ab, und es bedarf relativ weitgehender Vorannahmen auf den genannten Gebieten, um die Experimente überhaupt widerspruchsfrei durchführen zu können. Variiert man diese Vorannahmen, lösen sich die Probleme gelegentlich von selbst auf, oder werden im anderen Extrem paradox und damit unauflösbar.

Anhang A: Newcombs Problem

Ein übernatürliches Wesen, das sich meines Wissens noch nie bei Vorhersagen über die Zukunft geirrt hat, erklärt mir folgendes: „Ich stelle hier jetzt eine rote und eine blaue Kiste hin, und rühre sie dann nicht mehr an. Du darfst dir in einem Monat entweder beide oder nur die blaue nehmen. Wenn ich jetzt voraussage, dass du in einem Monat nur die blaue nehmen wirst, lege ich eine Million Euro hinein, andernfalls nichts. In die rote Kiste lege ich auf jeden Fall 1000 Euro.“

» Literatur: Nozick, Robert, 1969: "Newcomb's Problem and Two principles of Choice".

Anhang B: Gifträtsel

Wenn es mir heute abend gelingt, die Absicht zu haben, morgen mittag einen Becher Gift zu trinken, bekomme ich morgen früh eine Million Euro, wobei es keine Rolle spielt, ob ich den Becher am nächsten Tag austrinke oder nicht. Das Gift wirkt für 24 Stunden sehr unangenehm, hat aber keine Langzeitfolgen. Meine Absicht wird heute abend mit einem neuartigen Messgerät überprüft, auf dessen Zuverlässigkeit ich vertraue. Zusatzabsprachen mit Dritten sind nicht erlaubt.

» Literatur: Kavka, Gregory, 1983: "The Toxin Puzzle".