Newcombs Paradoxie – und einige ihrer verfehlten Lösungsversuche

Bastian FISCHER[1]
Universität des Saarlandes

I. Einleitung

Newcombs Paradoxie ist zweifelsohne eines der verzwicktesten Entscheidungsdilemmata und zugleich eines, für das noch keine Standardlösung innerhalb der Entscheidungstheorie existiert. Das Dilemma ist, anders als etwa Mackie (vgl. 1977, 216) glaubte und wie schließlich Frydman, O'Driscoll & Schotter (1982) nachwiesen, durchaus in der aktualen Welt anzutreffen, so etwa in Gestalt bestimmter Versionen der in der Volkswirtschaftswissenschaft analysierten, so genannten Zeit-Inkonsistenz-Probleme. Die Paradoxie ist zudem potentiell aufschlussreich für wissenschaftstheoretische Grundlagenfragen zur Verursachung und Determiniertheit und in diesem Rahmen insbesondere Aufhängepunkt für den Widerstreit zwischen Vertretern einer kausalen Entscheidungstheorie einerseits, der zufolge rationale Handlungen um dessen willen, was sie *verursachen*, ausgeführt werden, und einer evidentiellen Entscheidungstheorie andererseits, der zufolge rationale Handlungen – eventuell bloß – um dessen willen getan werden, wofür sie *Belege* liefern (vgl. Horwich 1985, 446).

II. Das Problem

Stellen wir uns ein Wesen vor, das bisher menschliche Handlungen sehr gut vorhersagen konnte. Dieses Wesen könnte ein Wahrsager, Psychologe, Neurowissenschaftler oder dergleichen sein. Seine Vorhersagen waren bisher immer richtig. In einem Raum stehen nun zwei Boxen Box A und Box B. Wir haben die Wahl zwischen zwei Handlungen: (1) Nehmen, was sich in beiden Boxen befindet ('***BEIDE***'), (2) *nur* nehmen, was sich in Box B befindet ('***EINE***'). Box A enthält in

[1] Bastian Fischer, Universität des Saarlandes, Philosophisches Institut, Postfach 15 11 50, D-66041 Saarbrücken, Germany; E-mail: bastian.fischer@mx.uni-saarland.de oder anlagezugross@gmx.net.

jedem Fall 1.000$, Box B entweder 0$ oder 1.000.000$. Der Inhalt von Box B wird gemäß der Vorhersage, die das Wesen trifft, bestimmt: Sagt das Wesen voraus, dass wir ***EINE*** wählen, so wird in Box B ein 1 Mio.$-Scheck gelegt, sagt es jedoch voraus, dass wir ***BEIDE*** wählen, so bleibt die Box B leer. Wir möchten die Summe des zu erhaltenden Geldes maximieren und rational handeln. Welche Wahl ist zu bevorzugen?

(i) Versuchen wir, für die Wahl ***EINE*** zu argumentieren. Angenommen, wir nehmen ***BEIDE***, so hat das Wesen dies mit hoher Sicherheit vorausgesehen, Box B wird also mit hoher Sicherheit leer sein. Dies bedeutet, wir erhalten bei ***BEIDE*** mit hoher Sicherheit am Ende nur 1.000$. Angenommen, wir nehmen ***EINE***, so wird das Wesen dies ebenfalls mit hoher Sicherheit vorausgesehen haben, Box B enthält also mit hoher Sicherheit 1 Mio.$, die wir auch erhalten. Daher sollten wir die Handlung ***EINE*** der Handlung ***BEIDE*** vorziehen.

(ii) Andererseits ist klar, dass das Wesen seine Vorhersage getroffen hat, bevor wir unsere Entscheidung verkünden. Das heißt, in dem Moment, in welchem wir uns entweder für ***BEIDE*** oder ***EINE*** entscheiden, ist der Inhalt der Box B schon bestimmt und durch unsere bloße Wahl nicht mehr zu verändern. Zwei Fälle sind möglich.

Fall 1: Box B ist leer. Bei ***BEIDE*** erhalten wir dann 1.000$; bei ***EINE*** 0$.

Fall 2: Box B enthält 1 Mio.$. Bei ***BEIDE*** erhalten wir dann 1.001.000$; bei ***EINE*** genau 1 Mio.$.

Wir sehen, dass wir in beiden möglichen Fällen, wenn wir ***BEIDE*** wählen, 1.000$ mehr erhalten, als wenn wir ***EINE*** wählen. Daher sollten wir die Handlung ***BEIDE*** der Handlung ***EINE*** vorziehen.

Wir haben hier für ein und dieselbe Fragestellung zwei Argumentationen als Antwortversuche vor uns, welche trotz jeweils scheinbar höchst akzeptabler Prämissen zu verschiedenen, sich gegenseitig ausschließenden Konklusionen gelangen. In diesem Sinne liegt mit diesem Entscheidungsdilemma eine praktische *Paradoxie* vor.[2]

Die beiden Argumente (i) und (ii) lassen sich in ihrer Überzeugungskraft durch weitere Überlegungen bestärken (vgl. Nozick 1969, 115f.):

Zu (i): Angenommen, viele unserer Bekannten haben bereits an diesem Experiment teilgenommen. Sie haben sogar beide Argumente gekannt. Diejenigen, die sich für ***EINE*** entschieden haben, sind alle Millionäre, diejenigen, die sich für ***BEIDE*** entschieden haben, haben alle nur 1.000$ erhalten. Sollten wir dann nicht vernünftigerweise denken, dass wir, wenn wir uns für ***BEIDE*** entscheiden, auch nur

[2]Diese Formulierung des Newcomb-Problems basiert auf Nozick (1969, 114f.).

1.000$ bekommen werden? Durch diese Verstärkung wird das Argument des Induktionsschlusses verdeutlicht, das auch Schlesinger der Eine-Box-Strategie zugute hält (1974, 209 ff.).

Argument (i) basiert zumindest im *evidentiellen* Rahmen der Entscheidungstheorie auf dem Prinzip des maximal zu erwartenden Nutzens: Aus einer Menge möglicher Strategien sollte man diejenige wählen, die den zu erwartenden Nutzen maximiert. Eine Aufstellung der Erwartungsnutzen der beiden möglichen Entscheidungen ***EINE*** und ***BEIDE*** in einer Ungleichung ergibt nach Auflösen, dass es bei den angegebenen Auszahlungswerten bereits genügen würde, dass die Wahrscheinlichkeit einer korrekten Vorhersage des Wesens echt größer als 50,05 % ist, damit der Erwartungsnutzen von ***EINE*** echt größer ist als der Erwartungsnutzen von ***BEIDE***.

Zu (ii): Der Zeitraum des Experiments wird etwas gedehnt, und ein unabhängiger Ratgeber darf gewinnmaximierende Entscheidungsanregungen geben. Wir nehmen an, dass das Wesen seine Vorhersage macht, und das Geld, je nach Vorhersage, entweder in Box B gelegt wird oder nicht, und das Wesen den Schauplatz verlässt. Es vergeht eine Woche. Wir gehen in den Raum mit den Boxen und sollen uns nun entscheiden, ob ***BEIDE*** oder ***EINE***. Das Wesen hat im vorliegenden Zustand keinen direkten Einfluss mehr auf den Inhalt der Box B. Entweder wartet die Million in Box B oder nicht. Wenn wir beide Boxen nehmen, haben wir in jedem Fall also 1.000$ mehr, als wenn wir nur nehmen, was in Box B ist. Unser nun erscheinender Ratgeber weiß vor unserer Entscheidung, was sich in den Boxen befindet. Entweder weiß er, (a) dass sich in Box A 1.000$ befinden (was wir ja auch wissen) und sich in Box B 1 Mio.$ befinden, oder er weiß, (b) dass sich in Box A 1.000$ befinden und sich in Box B nichts befindet. Es ist nicht schwer einzusehen, dass er uns raten würde, beide Boxen zu wählen, und das ganz gleich, ob er nun (a) oder (b) weiß.

Auch diese Verstärkung wird von Schlesinger übernommen. Aus der genannten gewinnmaximierenden Entscheidungsempfehlung des Beraters und der induktiven Verlässlichkeit des Wesens, welche die Eine-Box-Lösung stützt, möchte er sogar einen deduktiven Widerspruch ableiten, der die Zwei-Boxen-Strategie seines Erachtens als die profitablere plausibel macht und die Eine-Box-Strategie *ad absurdum* führt (211 ff.). Unseres Erachtens läuft dieser Ansatz für eine Verteidigung der Zwei-Boxen-Strategie allerdings darauf hinaus, lediglich die beiden *prima facie* zumindest gleich mächtigen Hörner des Entscheidungsdilemmas gegeneinander zu wetzen, und er ist

argumentativ letztlich ebenso wenig substantiell wie die Behauptung, der Materialismus sei falsch, weil er dem Dualismus widerspreche.

Argument (ii) für das Duoboxen basiert jedenfalls auf dem Dominanzprinzip: Wahlen können je nach Zuständen oder Ereignissen in der Welt verschiedene Nutzen haben. Eine Wahl *G* ist einer Wahl *H* genau dann vorzuziehen, wenn sie Wahl *H schwach dominiert*, wobei Wahl *G* die Wahl *H* schwach dominiert gdw. für jeden möglichen Weltzustand bzw. jedes Weltereignis gilt, dass der Nutzen der Wahl *G* mindestens genauso hoch ist wie der Nutzen der Wahl *H* und dass er für mindestens einen möglichen Weltzustand oder ein mögliches Ereignis echt höher ist. In Newcombs Problem dominiert ***BEIDE*** offensichtlich ***EINE***, und dies sogar stark, denn ganz gleich, ob die zweite Box mit der Million befüllt ist oder nicht, ***BEIDE*** zu wählen bringt 1.000$ mehr ein, als bei dem eben jeweilig gegebenen Füllzustand nur die eine Box zu nehmen.

III. Substantielle Lösungsansätze

Viele der bisher für Newcombs Problem angebotenen Lösungen sind, genau betrachtet, entweder fehlerhaft oder mindestens insofern mit einer bösartigen Zirkularität behaftet, als an entscheidenden Stellen zumindest implizit vorausgesetzt wird, dass – bei Monobox-Lösungen – das Wesen mit notwendiger Sicherheit die Handlungen des Akteurs richtig vorhersagt oder dass – bei Duobox-Lösungen – das Wesen die Handlungen eben nicht mit notwendiger Sicherheit korrekt vorhersagt. Die viel ersprießlichere Frage, ob eben empirisch nach allem, was wir wissen, das eine oder das andere dieser disjunktiven Dichotomie plausibel ist, wird jedenfalls in der Diskussion der Paradoxie selbst selten überhaupt auch nur angesprochen. So schließt Bach (1987), ein Monoboxer, etwa von vornherein explizit aus, dass man als Akteur seinen gesamten psychologischen Zustand kennen kann, auf welchem die Vorhersage des Wesens etwa beruhen könnte:

> "Since you have no way of knowing what your total psychological state was or how it provided PR[edictor] with a basis for his prediction, your only evidence for what he predicted is your actual choice, but you have not yet made that choice." (420)

Ledwig (2000), eine Duoboxerin, betrachtet die Newcombsche Entscheidungssituation dagegen als ein "Spiel gegen die Natur", bei welchem nach

ihrer Definition eben dieses Terminus gilt, dass man als Akteur mit einem endgültig festgelegten Zustand der Welt konfrontiert ist, nicht mit einer durch die eigene Entscheidung modifizierbaren und vorher ausgeführten Handlungsstrategie eines zweiten Akteurs; die Vorhersage des Wesens wird also als reines Naturereignis behandelt (vgl. 45, 279f.). Diese Betrachtungsweise impliziert, dass die Entscheidung des Akteurs, wenn auch nicht probabilistisch, so doch kausal unabhängig ist von dem jeweiligen Füllzustand der Boxen.

Ledwigs Begründung für das Einnehmen dieser Betrachtungsweise ist allerdings mehr schlecht als recht: Erstens scheint es im Wesentlichen auf einer Gegenüberstellung zwischen einem einfach gespielten Newcomb-Problen und einem mehrfach iterierten zu beruhen, wobei es sich im iterierten empfehle, sich in einigen Spielen dem Wesen gegenüber einen Ruf als Monoboxer zu verschaffen, damit es einen als Akteur entsprechend einschätzt, während sich einen solchen Ruf durch vorherige Aktionen zu verschaffen im einfachen Newcomb-Spiel als Option evidenterweise nicht in Frage komme (vgl. 280f.). Hiermit soll schon gezeigt sein, dass das iterierte Spiel einen zweiten genuinen Akteur beinhaltet, den das Wesen darstellt und den man durch seine Handlungsmuster beeinflusst, während die genuin akteursmäßige Natur des Wesens im einfachen Spiel überflüssig sei, weshalb die Betrachtung des letzteren als Spiel gegen eine nicht bewusst entscheidende Natur schon angemessen sei. Diese Gegenüberstellung macht zwar durchaus plausibel, dass man es im einfachen Spiel mit einem natürlichen Weltzustand zu tun hat, den man durch seine Aktionen nicht mehr verändern kann bzw. der von den eigenen Aktionen kausal unabhängig ist, sie kann aber kein schlüssiger Beweis dafür sein. Allein die vom Akteur im Vorlauf des Spiels gewählten Handlungsmuster, andere psychologische, eventuell auch mystisch-paranormale Parameter oder sein Gehirn könnten zumindest *prima facie* die Vorhersage des Wesens bestimmen, die dann eben nicht mehr als – in Ledwigs leicht anthropozentrisch-dualistischem Sinne – reines Naturereignis, sondern als Verstandesleistung des Wesens zu deuten wäre. Alternativ hierzu könnte man mit ungefähr gleichem Recht auch sagen, dass, statt dass sowohl Vorhersagewesen als auch Entscheider anthropische Akteursqualitäten zuerkannt bekommen, schlicht *beider* Handlungen als "reines Naturereignis" in Ledwigs Sinne erscheinen könnten: Mich als Akteur determiniert beispielsweise mein Gehirn in meiner Entscheidung, das Wesen misst die relevanten Parameter in meinem Kopf und macht davon ausgehend ebenso mechanisch seine Vorhersage, wie ich mich entscheide.

Darüber hinaus schließt Ledwig ohne substantielle Begründung das Bestehen von Rückwärtskausalität aus (vgl. 281). Eine solche merkwürdige Art der Kausalität würde freilich – selbst unter der Annahme der Adäquatheit einer strikt *kausalen* Entscheidungstheorie – das Monoboxen stützen, denn unter dieser Bedingung bestimmt man als Akteur nachträglich durch seine Wahl, was seit der Füllung schon immer in der zweiten Box ist.

Schließlich setzt die Betrachtungsweise Ledwigs, wie bereits angedeutet, voraus, dass Entscheidungsfreiheit besteht. Ihr Erklärungsmodell und ihre Empfehlung, beide Boxen zu nehmen, sind nicht mehr anwendbar, sollte sich herausstellen, dass ohnehin jeder mögliche Akteur in seinen Entscheidungen determiniert ist und eben zumindest letztlich keine aktive Rolle als Urstifter einer Kausalkette spielt. Ihre, zugegeben, gemäßigte Anforderung an einen freien Entscheidungswillen in dem Sinne, dass "[t]he decision maker can rule out the possibility that he attempts to carry out his decision, but finds himself unable to do so" (281), könnte auch in deterministischen Szenarien erfüllt sein, in welchen es (a) für den relevanten Fall keine determinierten Hürden gibt und auch (b) nie jemand die Widerspenstigkeit besitzt, gegen seine natürliche Basis zu entscheiden, sondern einfach so handelt, wie es sein Gehirn, seine Psyche oder – meinetwegen auch "anything goes" – etwa sein Horoskop anordnet.

Schon Hubin & Ross (1985, 444 ff.) vereinen diese beiden Ansätze von Bach und Ledwig. Sie argumentieren gewieft und formal sauber dafür, dass unter der Voraussetzung, dass der Newcomb-Vorhersager – im streng nomologisch-notwendigen Sinne – unfehlbar ist, der Akteur ganz klar nur die zweite Box nehmen sollte. Sie zeigen, dass bei Aufrechterhaltung der Entscheidungsfreiheit des Akteurs in dem Sinne, dass es sowohl möglich ist, dass er ***BEIDE*** wählt, als auch möglich, dass er ***EINE*** wählt, unter dieser Unfehlbarkeitsvoraussetzung allerdings die Bedingung, dass der Inhalt der zweiten Box von der Wahl des Akteurs unabhängig festgelegt ist, aufgegeben werden muss. Andererseits sollte der Akteur unter der Voraussetzung, dass seine Wahl eben doch keinen Einfluss auf den Boxeninhalt ausübt (und er frei entscheidet), ganz klar beide Boxen nehmen. Allerdings müsse dann wiederum die Bedingung der Unfehlbarkeit des Vorhersagers aufgegeben werden. Hubins & Ross' Urteil lautet daher: "The puzzle is overconstrained" (445). Sobald einmal die Rolle der angeführten Voraussetzungen und Bedingungen mit all ihren Implikationen verstanden sei, "the force of both the one-box and the two-box arguments can be appreciated. The

solution lies not in choosing between the two arguments but in choosing between different coherent formulations of Newcomb's problem" (446).

Schlussfolgernd sollte man auf der Suche nach einer Lösung für das Dilemma das Augenmerk wohl auf die Frage richten, welche Formulierung des Newcomb-Problems nach allem, was wir wissen, die empirisch kohärentere ist und ob eben das Wesen nun notwendigerweise richtig vorhersagt oder aber eine Widerspenstigkeit gegen seine Vorhersage möglich ist. Wenn letzteres wahr ist, so wird der Eine-Box-Lösung ein Großteil ihrer argumentativen Basis entzogen.

Bei Lenzen (1997) finden wir nun eine auf den ersten Blick raffinierte Ausdifferenzierung der verschiedenen Möglichkeiten bezüglich dieser gerade anvisierten Suchrichtung, die wiederum den Schluss nach sich zieht, dass man beide Boxen nehmen sollte. Was die Option eines unfehlbaren Vorhersagers betrifft, so scheint Lenzen geltend zu machen, dass die Konstanthaltung der Vorhersagesicherheit über verschiedene mögliche Welten hinweg darauf hinaus laufe zu behaupten, dass

> "(z entscheidet sich [mit hoher Sicherheit] für [***EINE***] gdw. W[esen] vorhergesagt hat, daß z [***EINE***] wählt) und
> (z entscheidet sich [mit hoher Sicherheit] für [***BEIDE***] gdw. W[esen] vorhergesagt hat, daß z [***BEIDE***] wählt)" (*) (171)

mit mindestens (natur-)*gesetzlicher* Notwendigkeit gilt, d.h. in allen Welten, wo nur die gleichen Naturgesetze wie in der unseren herrschen. Lenzen beruft sich hierbei auf Goodmans neues Rätsel der Induktion aus Goodman (1955), seit welchem klar sei, dass (*) nur unter dieser Notwendigkeitsbedingung den irrealen Konditionalsatz "wenn z sich nicht für ***EINE*** entschiede, dann hätte das Wesen auch nicht vorhergesagt, dass z ***EINE*** wählt" zu stützen vermag. Ihn führt dieser Gedankengang zu einer Analyse eines deterministischen Gesetzes, welches die Wahlen des Akteurs unausweichlich bestimmbar macht und auf dessen Basis das Newcomb-Problem als echtes *Entscheidungs*problem verschwinden, ja "jeglichen Sinn" verlieren würde (172).

Damit setzt aber auch Lenzen – wie Ledwig – für ernst zu nehmende Newcomb-Probleme einen Indeterminismus bezüglich menschlicher Handlungen voraus. Stellen Sie sich vor, es würde sich herausstellen, dass alle Handlungen ohnehin determiniert sind und der Kompatibilismus bezüglich der Willensfreiheit kohärent ist. Freilich würde irgendeine echte Entscheidungsempfehlung im traditionellen Sinne sich dann vielleicht erübrigen. Nichtsdestotrotz wäre es weiterhin zumindest interessant zu sehen, wer denn nun "das Match gewinnt": Monoboxer oder Duoboxer? Vielleicht könnte

man sogar doch – im Falle von zumindest vorläufiger *neurologischer* Determiniertheit – wenigstens hirnchirurgisch Einfluss nehmen, so dass bestimmte Menschen umgepolt werden können und dann determiniert so agieren wie die bisherigen Gewinner. Was wäre, wenn in diesem Fall die deterministisch handelnden Monoboxer erhobenen Hauptes das Newcomb-Spielfeld verlassen und sagen: "Ja, wir haben eben die bessere physikalische Basis, die uns so 'entscheiden' lässt, dass wir reich werden: wir haben richtig gewählt, ihr Duoboxer habt einen Körper, der euch nur zur schlechten Wahl führt, ihr könnt einfach (im wahrsten Sinne) nicht so wählen, dass ihr reich werdet...."? Würden Sie sich dann nicht auch umoperieren lassen, wenn Sie schnell viel Geld brauchten?

Es ist hierzu anzumerken, dass es mit Alberts & Heiners (2001) volkswirtschaftswissenschaftlicher Analyse des Newcomb-Problems bereits einen Verteidigungsansatz für das Monoboxen gibt, der einen Determinismus bzgl. menschlicher Handlungen schlicht postuliert: In diesem Rahmen sind die Spieler in einer Population mit unveränderlichen Motivationspaketen ausgestattet, die sie zwingend zu bestimmten Entscheidungen führen, und das Monoboxen wird einfach dadurch als die bessere Strategie ausgewiesen, dass die jeweiligen Motivationspakete und die zugehörigen Auszahlungen den Monoboxern gegenüber den Duoboxern einen evolutionären Vorteil verschaffen. An dieser Stelle erfordert eine Verteidigung des Duoboxens, wie etwa Lenzen oder Ledwig sie zu führen beabsichtigen, eine Widerlegung der Determiniertheit der Akteursentscheidung oder zumindest eine Widerlegung ihrer Unumkehrbarkeit bis zum letzen Moment des Entscheidungsprozesses. Diese liefern aber weder Lenzen noch Ledwig.

Lenzen urteilt weiterhin:

> "Der von manchen AutorInnen erteilte Ratschlag 'Glaube dem Vorhersager, nimm bescheiden nur die [zweite Box] und werde Millionär!' erweist sich deshalb als *nutzlos*, sofern er dem Handelnden zum Zeitpunkt der Entscheidung [...] zu einem 'rationalen' oder gewinnmaximierenden Entschluß verhelfen soll. Zu *diesem* Zeitpunkt sind die Würfel längst gefallen!" (172f., Lenzens Hervorhebungen)

Lenzen müsste hier aber die rückwärtige Kausalität noch ausschließen, um behaupten zu können, dass das Monoboxen in der Tat "nutzlos" sei. "Zu *diesem* Zeitpunkt sind die Würfel längst gefallen!" mag durchaus wahr sein, was aber, wenn sie trotzdem immer oder zumindest ausreichend oft so gefallen sind, wie es die jetzige

Entscheidung bei (zumindest ausreichend hoher) Vorhersagesicherheit aktiv bestimmen würde? Lenzen (persönliche Kommunikation) gesteht ein, dass sein Urteil in der Tat präsupponiert, dass hier keine Rückwärtskausalität im Spiel ist, allerdings erscheine ihm die Legitimität dieser stillschweigenden Voraussetzung nach wie vor "allzu evident". Wenn etwa Dummett (1964) meines Erachtens auch nicht völlig schlüssig gezeigt hat, wie sich das Bestehen rückwärtiger Kausalität positiv testen lässt, so hat er durchaus gezeigt, dass der Glaube an diese merkwürdige Art von Kausalität keinesfalls so absurd ist, wie er es vielen – eben auch Lenzen – anfänglich zu sein scheint. Lenzens intuitionsgestützter Ausschluss von Rückwärtskausalität erscheint daher, wenn auch nicht unplausibel, so doch etwas voreilig.

Für den Fall des nicht gänzlich unfehlbaren Vorhersagers entwirft Lenzen exemplarisch das folgende Ausgangsmodell des Newcomb-Szenarios. Es gab 100 Monoboxer und 100 Duoboxer in einer Spielreihe, und 99 Monoboxer bekamen die Million, ein Monoboxer ("Pechvogel") ging aber leer aus, und 99 Duoboxer bekamen nur 1.000$, ein Duoboxer ("Maximalgewinner") konnte aber 1.001.000$ abräumen. Ein solches experimentelles Ergebnis deutet Lenzen nun als Beleg dafür, dass es prinzipiell jedem Spieler *möglich* sei, die Vorhersage des Wesens zu vereiteln:

> "Bei der [...] Variante [des Problems mit] einer mit 99 % zwar sehr hohen, aber eben nicht absolut sicheren Zuverlässigkeit besteht keine Möglichkeit zu folgern, daß wenn irgendjemand sich anders entschieden hätte als er sich de facto entschieden hat, dann das mächtige Wesen dies vorhergesehen hätte: Der Fall des Pechvogels x* bzw. des Maximalgewinners y* zeigt dies aufs deutlichste" (170).

Wirklich? Lenzen geht hier scheinbar ohne weiteres von 'nicht absolut sicher richtiger Vorhersage' zu 'bloß kontingent richtiger Vorhersage' über. In einem gewissen frequentistisch-epistemischen wahrscheinlichkeitstheoretischen Rahmen, innerhalb dessen Wahrscheinlichkeiten schlicht relative Häufigkeiten der relevanten Ergebnisse sind, scheint dieser Schluss berechtigt zu sein. Es gibt allerdings dagegen konkurrierende wahrscheinlichkeitstheoretische Ansätze, denen zufolge Wahrscheinlichkeiten etwa schlicht reale Dispositionen von – eventuell auch pluralen – Entitäten sind (vgl. Hájek 1997). In einem solchen propensitätstheoretischen Rahmen wäre es durchaus zu rechtfertigen, dass das Wesen in allen Welten aus einer Gruppe von 100 Leuten genau 99 richtig einschätzt und eine Person falsch, und dies auf ähnliche Weise, wie es bei zwei verschränkten Quantenobjekten der Fall zu sein

scheint, dass genau eines davon in dem einen Zustand ist, das andere aber – bei gewissen Hintergrund- und Beobachtungsbedingungen notwendig – in einem anderen Zustand. Folglich ließe sich eben auch nicht ausschließen, dass, wenn irgendjemand der 99 sich anders entschieden hätte, als er sich *de facto* entschieden hat, dann das Wesen dies vorhergesehen hätte. Folglich spräche in einem propensitätstheoretischen Rahmen auch bei Varianten des Newcomb-Szenarios mit einer bloß 99%igen Verlässlichkeit des Wesens zunächst nichts gegen eine Verteidigung der Monobox-Strategie mit der Begründung, dass das Wesen eben hier *notwendigerweise* eine höhere Vorhersageverlässlichkeit habe als die erforderlichen 50,05 %. Lenzen setzt, grob gesagt, die Korrektheit einer Ausprägung frequentistischer Wahrscheinlichkeitstheorie voraus, der zufolge ich auch nicht hätte Lungenkrebs bekommen können, sobald auch nur ein Mensch lebenslang ohne Lungenkrebs existiert. Diese auch nicht gänzlich unplausible Theorie ist allerdings stark begründungsbedürftig.

Äußerst originell ist schließlich eine Theorie aus dem Monoboxer-Lager, die ein Modell für eine Realisierung des Newcomb-Szenarios mit angeblich rückwärtigen kausalen Abhängigkeiten zwischen Wesen und Akteur enthält: Schmidt (1998) stellt sich eine Sammlung kleinster Wesen aus "Tinionen" vor, welche die Entscheidung des Akteurs aus ihrer Perspektive so sicher vorhersagen können, wie wir die Bewegung der Planeten berechnen können. Der Haken daran ist, dass diese von uns noch nicht erforschten und sogar noch nicht erforschbaren Wesen zur Kommunikation mit uns einen Vorlauf von mindestens 24h 1min brauchen, aber trotzdem auf ihrer Ebene direkt einen für Menschen unlesbaren, in Banken jedoch konvertierbaren Minischeck über die Million oder eben null Geldeinheiten ausstellen können; leider können sie scharfe Vorhersagen aber nur in einer Zeitspanne von genau 24h machen, innerhalb deren letzter Minute wir uns als Akteure auch verbindlich entscheiden sollen. Es folgt angeblich, dass diese Wesen uns nicht einmal selbst mitteilen könnten, welche Vorhersage sie getroffen haben.

Dieser Ansatz Schmidts wurde von Ledwig (2001) kritisiert, die allerdings einen grundlegenden Fehler übersieht. Ob es sich bei der von Schmidt erwogenen Kausalbeziehung wirklich um eine – wie er es behauptet – zeitlich rückwärtige handelt, sei einmal dahingestellt; es ist jedoch, wie eine einfache Berechnung (siehe Anhang) zeigt, aufgrund der relativistischen Zeitdilatation zumindest physikalisch möglich, auch die Vorhersage von Schmidts Tinionenwesen zu vereiteln, weshalb auch dieser eine

notwendig richtige Vorhersage schlicht und einfach *postulierende* Lösungsansatz für Newcombs Problem verfehlt ist.

Mir scheint die Vorarbeit für eine solche Vereitelung der Vorhersagesicherheit des Vorhersagewesens – sprich: an eine Information über den Füllzustand der Box B zu kommen – auch in von Schmidts Szenario abweichenden Newcomb-Spielsituationen zumindest immer vorstellbar zu sein.[3] Wenn nun etwa Chalmers' (2002) These stimmt, dass zumindest viele – hier jedenfalls relevante – Arten von Vorstellbarkeit Möglichkeit implizieren, so scheint mir dann mit der Möglichkeit der Informationsgewinnung über den Füllzustand von Box B auch die Vereitelung der Vorhersagekorrektheit selbst immer möglich zu sein, zumindest für Trotzwesen, d.h. Entitäten, welche bloß einen primitiven Trotzalgorithmus der Form "Wenn nur A, B zur Wahl stehen, so wähle A gdw. jemand meint, du würdest B wählen" implementiert haben. Somit scheint es mir unter diesen Prämissen sowie gemäß Hubins & Ross' obiger Ausdifferenzierung des Newcomb-Dilemmas für alle Trotzwesen rational zu sein beide Boxen an sich zu nehmen.

Anhang: *Berechnung zu Schmidts Modell*

Es stehe ein Raumschiff bereit, mit welchem Schmidt samt der ersten Box mit dem makroskopischen 1.000$-Scheck und den sie umgebenden und eventuell für die Vernichtung des 1.000$-Schecks bei Vorhersage von ***EINE*** sorgenden[4] Tinionen in dem Moment startet, in welchem er die Nachricht der Tinionen über die Spielregeln ihres Newcomb-Spiels zu Ende gelesen hat. Im selben Moment lege ich die zweite Box mit dem Minischeck (die Bank ist gleich nebenan) in den Detektorraum, in dem Tinionen ihn in 24h 1min umwandeln können. Schmidt beschleunigt geradlinig konstant über einen Zeitraum t mit einer Beschleunigung a. Nach t ändert er die Beschleunigung auf 0 und hält sie für $t_1 = (24\ h + 1\ min)/2 – 2t$ konstant so. Nach diesem Zeitraum bremst er mit -a Beschleunigung t lang ab bis zum Stillstand, behält diese Beschleunigung aber nochmals t lang bei und beginnt so zurückzureisen. Dann reist er wieder $t_1 = (24\ h + 1\ min)/2 – 2t$ lang mit gleich bleibender Geschwindigkeit, wonach er t lang mit Beschleunigung a abbremst, so dass er auf Erden wieder sanft zum Stillstand kommt. Die Zeit, die während dieser Reise auf der Erde vergeht, beträgt offensichtlich 24h 1min. Wenn er zurückkommt, ist die Detektion des Minischecks also gerade fertig. Die Zeit, die diese Reise ihn selbst und die Box, die Tinionen und den

[3] Eine Prämisse, deren Wahrheitsstatus ich hier leider nicht mehr diskutieren kann; die Kritik an anderen Modellen war mir wichtiger.

[4] Dies garantiert in Schmidts Modell, dass man nicht nachträglich, d.h. nachdem man sich einmal verbindlich für nur die eine Box entschieden hat, doch noch die zweite Nutzen bringend an sich reißen kann. Den 1.000$-Scheck lassen die Tinionen also kurz nach der Entscheidung des Akteurs in Flammen aufgehen, wenn er nur die eine Box genommen hat, er bleibt bei ***BEIDE*** intakt. Man sollte anmerken, dass die Tinionenwesen mit Schmidt nur brieflich kommunizieren und nicht unmittelbar handeln können, wie es Menschen könnten, indem sie einfach die erste Box nach der Entscheidung wegstellen würden.

schon großen 1.000$-Scheck kostet, ist – zumindest bei Geltung der Zeitdilatation für Tinionen – gegeben durch

$$t_{Raumschiff} = 4\int_0^t \sqrt{1-\frac{(a\cdot t)^2}{c^2}}\,dt + (24h+1min-4t)\sqrt{1-\frac{(a\cdot t)^2}{c^2}} =$$

$$2\cdot\left[\frac{a\cdot t\cdot\sqrt{1-\frac{(a\cdot t)^2}{c^2}}+c\cdot\arcsin\left(\frac{a\cdot t}{c}\right)}{a}\right]_0^t + (24h+1min-4t)\sqrt{1-\frac{(a\cdot t)^2}{c^2}},$$

womit wir bei t = 18s und a = 16.666.666,6666m/s² als $t_{Raumschiff}$ = 56,793010615588294s hätten. Die 3,206989384411706s, die Schmidt von seiner Entscheidungszeit dann noch bleiben, reichen vielleicht nicht mehr ganz aus, um vor dem endgültigen, letztlich entscheidenden Ansichnehmen der Boxen den Frühstückskaffee auszutrinken, aber da er mittlerweile ohnehin kalt ist, macht das nichts, sie reichen aber durchaus dazu aus, dass ich Schmidt ganz rasch sage, was auf dem umgewandelten Scheck steht. Steht da "0", muss Schmidt bloß die weitgereiste Box mit dem 1.000$-Scheck schnell auf den Tisch zurückstellen und nur die zweite an sich nehmen, steht da "1.000.000$", muss Schmidt nur rasch beide Boxen an sich nehmen. Die Vorhersagesicherheit der Tinionenwesen ist damit widerlegt. Als geringfügige Schwierigkeit dürfte sich Schmidts Überleben bei einer solchen Beschleunigung erweisen; die menschliche Robustheit halten wir jedoch für eine kontingente Angelegenheit, über die wir uns hinwegsetzen dürfen, wenn Schmidt einfach von der Existenz sehr merkwürdiger Teilchen ausgehen darf. Auch ihm geht es lediglich um die Vorstellbarkeit des Szenarios im Rahmen der physikalischen Gesetze. In diesem Fall sollte man gleiche Maßstäbe anlegen dürfen, auch wenn dieser Widerlegungsversuch dann möglicherweise in den Verruf gerät, bloß auf einem *ad hominem tu quoque*-Fehlschluss zu beruhen. Vielleicht muss Schmidt selbst aber auch gar nicht mitfliegen, und es reicht, bloß die Box mit dem Scheck und den Tinionen wegzuschicken.

Wir schließen, dass die Behauptung einer hier vorliegenden *notwendig* geltenden Kausalitätsrelation sehr wahrscheinlich fehlerhaft ist. Auch wenn mein Einwand nicht haltbar sein sollte, bleibt zu sagen, dass das Tinionenszenario, ebenso wie etwa Bachs Ansatz, viel zu *ad hoc* auf die Unmöglichkeit einer Widerlegung der Vorhersagesicherheit der Tinionenwesen getrimmt ist, um Schmidt vor dem Vorwurf der Zirkularität zu bewahren.

Danksagung

Ich danke Helge Rückert, der mich auf dieses Problem aufmerksam gemacht hat, und an dieser Stelle unnennbar vielen anderen für Diskussionen.

Literatur

Albert, M. und Heiner, R. A. (2001). "An Indirect-Evolution Approach to Newcomb's Problem." *Center for the Study of Law and Economics Discussion Paper* 2001-01, Saarbrücken, 2001. Abdruck online (13. Feb. 2007) <http://www.uni-saarland.de/fak1/fr12/csle/publications/2001-01_newc.pdf>. Auch als: Albert, M. und Heiner, R. A. (2003). "An Indirect-Evolution Approach to Newcomb's Problem." *Homo Oeconomicus* 20: 161-194.

Bach, K. (1987). "Newcomb's Problem: The $ 1,000,000 Solution." *Canadian Journal of Philosophy* 17: 409-426.

Broome, J. (1989). "An Economic Newcomb Problem." *Analysis* 49: 220-222.

Chalmers, D. (2002). "Does Conceivability Entail Possibility?" in: Gendler, T. und Hawthorne, J. (Hrsg.). *Conceivability and Possibility*. Oxford University Press. 145-200. Online-Abdruck (11. Feb. 2007) <http://consc.net/papers/conceivability.html>.

Dummett, M. (1964). "Bringing about the Past." *Philosophical Review* 73: 338-359.

Frydman, R., O'Driscoll, G. P. und Schotter, A. (1982). "Rational Expectations of Government Policy: an Application of Newcomb's Problem." *Southern Economic Journal* 49: 311-319.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Hájek, A. (1997). "Fifteen Arguments Against Finite Frequentism." *Erkenntnis* 45: 209-227.

Horwich, P. (1985). "Decision Theory in Light of Newcomb's Problem." *Philosophy of Science* 52: 431-450.

Hubin, D., und Ross, G. (1985). "Newcomb's Perfect Predictor." *Noûs* 19: 439-446.

Ledwig, M. (2000). *Newcomb's Problem*. University of Konstanz. (dissertation) http://www.ub.uni-konstanz.de/kops/volltexte/2000/524/.

Ledwig, M. (2001). "Newcomb's Problem and Backwards Causation." in: W. Spohn, M. Ledwig und M. Esfeld (Hrsg.). *Current Issues in Causation*. Paderborn: Mentis. 135-149.

Lenzen, W. (1997). "Die Newcomb-Paradoxie – und ihre Lösung." in: W. Lenzen (Hrsg.). *Das weite Spektrum der analytischen Philosophie: Festschrift für Franz von Kutschera*. Berlin und New York: de Gruyter. 160-177.

Mackie, J. L. (1977). "Newcomb's Paradox and the Direction of Causation." *Canadian Journal of Philosophy* 7: 213-225.

Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice." in: N. Rescher, D. Davidson und C. G. Hempel (Hrsg.). *Essays in Honor of Carl G. Hempel.* Dordrecht: Reidel. 114-146.

Schlesinger, G. (1974). "The Unpredictability of Free Choices." *British Journal for the Philosophy of Science* 25: 209-221.

Schmidt, J. H. (1998). "Newcomb's Paradox Realized with Backward Causation." *British Journal for the Philosophy of Science* 49: 67-87.