Universität Konstanz

Geisteswissenschaftliche Sektion

Fachbereich Philosophie

Proseminar: Entscheidung und Spiele

Dozent: Luc Bovens

WS 02/03

Thema:

Das Newcombs Problem

- Eine rationale Lösung -

Christian Hartz

Matrikel Nr. 01/494616

Semesteranschrift:

Kindlebildstr. 7a

78467 Konstanz

Telefon: 07531/698223

Email: christian.hartz@christian-hartz.de

http://www.christian-hartz.de

Abgabetermin: 31.03.2003

Inhaltsverzeichnis

Kapitel 1

Einleitung

Ein entscheidungstheoretisches Grundlagenproblem innerhalb der praktischen Philosophie stellt das Newcombs Problem (Nozick, 1969) dar.
Es gehört zu einer Reihe von Problemen und Paradoxa, die dazu genutzt werden praktische Probleme des wirklichen Lebens, wie beispielsweise der Ökonometrie, Politik oder auch der Theologie, entscheidungstheoretisch darzustellen und zu erklären.[1]
Die Bedeutung des Newcombs Problems ergibt sich daraus, daß zu seiner Lösung zwei gegenteilige Anschauungen normativer Rationalität angewandt werden können, die zu unterschiedlichen Lösungen des Problems führen. Durch diese nicht eindeutige Lösbarkeit kommt es zu weiteren Fragen, die beispielsweise die Kausalität oder die Vorhersagbarkeit von Problemstellungen und Entscheidungen betreffen. Diese Probleme und Fragen sind es, die das Newcombs Problem für die Entscheidungstheorie so wichtig gemacht haben.
Der Versuch eine eindeutige Lösung für das Newcombs Problem zu finden, spaltet die Entscheidungtheoretiker in zwei Lager: auf der einen Seite die "1 Boxer", von denen David Gauthier ein Vertreter ist, und auf der anderen Seite die "2 Boxer", zu denen R. Eric Barnes gezählt werden kann.[2] Selbst innerhalb der beiden Lager werden unterschiedliche Argumentationen verwendet, um die Lösung des Problems zu erreichen. Ich werde im Folgenden die verschiedenen

[1] Andere prominente Bespiele sind unter anderem: das Toxin Puzzle (Kafka, 1983), Parfits hitchtaker example (Parfit, 1984), das Prisoners dilemma, Gideons Paradox, Fishers Problem und Simpsons Paradox.

[2] Die Bezeichnung der beiden Lager ist zu diesem Zeitpunkt natürlich noch nicht ersichtlich, wird aber im folgenden deutlich gemacht.

Argumentationen der beiden Lager gegenüberstellen.

Dazu werde ich im Folgenden zunächst das Newcombs Problem kurz darstellen, um danach den traditionellen Weg zur Lösung des Problems zu erörtern. Daran anschließend werde ich einen entscheidenden Teil des Newcombs Problems, den Predictor, genauer betrachten und auf seine Relevanz zur Findung einer Lösung des Problems, im Rahmen des traditionellen Ansatzes, eingehen. Des weiteren werde ich diskutieren, wie mit der Theorie des rationalen Plans versucht wird, das Newcombs Problem zu lösen. Abschließend zeige ich noch Parfits Problem als Analogie zum Newcombs Problem.

Kapitel 2

Das Newcombs Problem

2.1 Problemstellung

Das Newcombs Problem wurde von Nozick (1969) aufgeworfen und läßt sich wie folgt beschreiben:
Man wird in einen Raum gebracht in dem sich zwei Boxen befinden. Die erste Box ist durchsichtig und es befinden sich stets 1000 Dollar in ihr. In der zweiten Box, die undurchsichtig ist, können sich entweder 1 Million Dollar oder aber null Dollar befinden. Dann verläßt man den Raum wieder und wartet darauf, daß man ein zweites Mal in den Raum gebracht wird, um sich für eine von zwei Handlungsalternativen zu entscheiden: (1) Man kann sich nur für Box 2 entscheiden ("1 Boxer"), oder aber (2) man kann sich dazu entscheiden, beide Boxen zu nehmen ("2 Boxer"). Die Entscheidung darüber, wieviel Dollar sich in der zweiten Box befinden, trifft ein Predictor in der Zeit, in der man sich nicht im Raum befindet und noch keine Entscheidung abgegeben hat. Dieser Predictor ist allwissend und daher in der Lage, mit großer Sicherheit (Vorhersagegenauigkeit p) vorherzusagen, wie man sich entscheiden wird. In Abhängigkeit von seiner Vorhersage legt der Predictor dann entweder die 1 Million Dollar (Vorhersage: Man entscheidet sich nur für Box 2) oder aber null Dollar (Vorhersage: Man entscheidet sich für beide Boxen) in die zweite Box. Nachdem der Predictor entweder 1 Million Dollar oder null Dollar in die zweite Box gelegt hat, wird man wieder in den Raum zurückgebracht und gibt dann seine Entscheidung bekannt.

Um sich das Problem noch einmal zu verdeutlichen kann man es in Form einer Matrix darstellen (siehe Tabelle 2.1).

Tabelle 2.1: Das Newcomb Problem

	Box 2 leer	Box 2 nicht leer
"1 Boxer"	0 $ $1-p$	1M $ p
"2 Boxer"	1000 $ p	1M + 1000 $ $1-p$

Im Folgenden stelle ich unterschiedliche Herangehensweisen an das geschilderte Problem dar, die zu den zwei in der Einleitung genannten Lösungen führen: die "1 Boxer" Lösung und die "2 Boxer" Lösung.

2.2 Der traditionelle Ansatz

Traditionell wird bei diesem Problem der "1 Boxer" als rationale Lösung angesehen. Zu dieser Entscheidung gelangt man, indem man für jede der Handlungsalternativen die erwartete Auszahlungssumme dieser Alternative berechnet. Die erwartete Auszahlungssumme einer Handlungsalternative ergibt sich, indem man jeden möglichen Ausgang der Handlungsalternative mit der entsprechenden Wahrscheinlichkeit gewichtet.[1]
In unserem Fall ergibt sich mit p als Vorhersagegenauigkeit des Predictors für die beiden Handlungsalternativen die allgemeine Gleichung:

$$\begin{aligned} \text{1 Boxer} &= 0 * (1-p) + 1000000 * p \\ \text{2 Boxer} &= 1000 * p + 1001000 * (1-p) \\ &= 1000 * p + 1001000 - 1001000 * p \end{aligned} \tag{2.1}$$

[1]Bei dieser Art der Betrachtung spielt also nur der Erwartungswertes der Auszahlungssumme eine Rolle in der Nutzenfunktion. D.h. wir setzen in diesem Fall Risikoneutralität des entscheidenden Individuums voraus.

Nimmt man nun eine hohe Vorhersagegenauigkeit an und setzt für p beispielsweise den Wert 0,9 (also eine Vorhersagegenauigkeit des Predictors von 90 Prozent) ein, so erhalten wir das folgende Ergebnis:

$$\begin{aligned} \text{1 Boxer} &= 0 * (1 - 0,9) + 1000000 * 0,9 \\ &= 900000 \\ \text{2 Boxer} &= 1000 * 0,9 + 1001000 * (1 - 0,9) \\ &= 900 + 100100 \\ &= 101000 \end{aligned} \tag{2.2}$$

Man sieht, daß es für die angenomme Wahrscheinlichkeit von 90 Prozent, bei dieser Art der Betrachtung des Problems am sinnvollsten ist, sich nur für eine Box zu entscheiden, da hier die erwartete Auszahlungssumme größer ist, als bei der 2 Boxer Strategie.

Man kann nun für diese Herangehensweise die Grenzwahrscheinlichkeit bestimmen, für die die "1 Boxer" Entscheidung gerade noch besser ist als die "2 Boxer" Entscheidung. Dazu werden die beiden Gleichungen der Handlungsalternativen in einer Ungleichung gegenübergestellt und nach der Wahrscheinlichkeit, p, aufgelöst:

$$\begin{aligned} 1000000 * p &> 1000 * p + 1001000 * (1 - p) \\ 2000000 * p &> 1001000 \\ p &> 0,5005 \end{aligned} \tag{2.3}$$

Die "1 Boxer" Entscheidung ist also für alle Wahrscheinlichkeiten $p > 0,5005$ der "2 Boxer" Entscheidung überlegen. Das heißt, daß die "1 Boxer" Lösung immer dann bei Anwendung des traditionellen Ansatzes gewählt wird, wenn der Predictor bessere Vorhersagen liefert, als das einfache Raten, das eine Vorhersagewahrscheinlichkeit von 50 Prozent hätte.

Was bedeutet aber überhaupt die Wahrscheinlichkeit p in diesem Problem? p ist die Wahrscheinlichkeit, mit der der Predictor die Entscheidung des Handeln-den vorhersagen kann. Können wir aber wirklich von einer Treffergenauigkeit des Predictors von mehr als 50% ausgehen? Schauen wir uns hierzu mal den Predictor an.

2.3 Der Predictor

2.3.1 Vorhersagewahrscheinlichkeit

Der Predictor wird bei der traditionellen Lösung des Problems als etwas Allwissendes dargestellt, das genau verhersagen kann, wie man sich entscheidet. Diese Allwissenheit begründet dann eine entsprechend hohe Trefferwahrscheinlichkeit für die Richtigkeit der Vorhersagen des Predictors. Dafür müßte der Predictor allerdings jede Person genau kennen um zu wissen, wie diese sich entscheidet oder er müßte eine allwissende Maschine sein, die in die Köpfe der Menschen schauen kann und deren Gedanken liest. Kann ein Predictor dies?

Tabelle 2.2: Das Newcomb Problem 50%

	Box 2 leer	Box 2 nicht leer
"1 Boxer"	0 \$ $0,5$	1M \$ $0,5$
"2 Boxer"	1000 \$ $0,5$	1M + 1000 \$ $0,5$

Dies wäre nur möglich, wenn es sich bei diesem Predictor um etwas göttlich Allwissendes handelt, da nur dann die Eigenschaft besteht, genau zu wissen wie jemand entscheidet. Es handelt sich aber um eine vom Menschen programmierte Maschine und kann daher die Eigenschaft der Allwissenheit nicht besitzen.

Da man nun aber nicht von einer Allwissenheit ausgehen kann, ist es auch nicht mehr richtig eine hohe Trefferwahrscheinlichkeit für die Vorhersagen des Predictors zu verwenden. Wir müssen vielmehr mit einer Wahrscheinlichkeit

von 50 Prozent arbeiten (siehe Tabelle 2.2), da die Vorhersagen des Predictors eigentlich nur durch eine Art raten zustande kommen. Wenn wir nun aber mit dieser Wahrscheinlichkeit von 50 Prozent rechnen, erhalten wir eine andere Lösung für das Problem.
Setzen wir also die Wahrscheinlichkeit p von 0,5 mal in die Ursprungsgleichung 2.1 ein, so erhalten wir als Ergebnis, daß es besser wäre sich für beide Boxen zu entscheiden:

$$\begin{aligned} \text{1 Boxer} &= 0 * (1 - 0,5) + 1000000 * 0,5 \\ &= 500000 \\ \text{2 Boxer} &= 1000 * 0,5 + 1001000 * (1 - 0,5) \\ &= 500 + 500500 \\ &= 501000 \end{aligned} \tag{2.4}$$

Man sieht also, daß die Genauigkeit der Vorhersagen des Predictors eine entscheidende Rolle bei der Lösung des Problems auf traditionellem Weg spielt.

2.3.2 Relevanz der Vorhersagewahrscheinlichkeit

Zu dem selben Ergebnis wie bei der Betrachtung der Genauigkeit der Vorhersagen des Predictors gelangt man auch, wenn man sich das folgende überlegt: Der Predictor trifft seine Entscheidung bevor man selber seine Entscheidung trifft, denn wenn man den Raum das zweite mal betritt, dann sind in der zweiten Box entweder 1 Million oder aber kein einziger Dollar, abhängig von der Vorhersage des Predictors. Die Entscheidung des Predictors ist aber bereits gefallen und somit kann die Entscheidung, ob man eine oder beide Boxen nimmt, keinen Einfluß mehr auf das Handeln des Predictors haben, da dieser bereits gehandelt hat. Der Entscheider hat ja auf alle Fälle die Möglichkeit seine getroffene Entscheidung im letzten Moment zu ändern. Für den Entscheider ergibt sich also nur das Problem, daß er, bei betreten des Raumes, nicht weiß, ob 1 Million oder 0 Dollar in der zweiten Box sind. Er kann allerdings überhaupt nichts darüber aussagen, wie wahrscheinlich die beiden Möglichkeiten sind, so daß er jeder Möglichkeit die gleiche Wahrscheinlichkeit von $p = 0,5$ zurechnen muß. Selbst wenn der Predictor allwissend wäre, so ist es der Entscheider nicht, so daß er nicht weiß, welche Entscheidung der Predictor getroffen hat. Der Entscheidende weiß, daß die Vorhersage kausal unabhängig von seiner eigenen Entscheidung ist.

2.4 Der Rationale Plan

Man gelangt ebenfalls zu der "2 Boxer" Lösung, wenn man sich kritisch mit den Argumenten von Gauthier (1989) auseinandersetzt, der ein Verfechter der "1 Boxer" Lösung ist.
Gauthier geht davon aus, daß jede Handlung dann rational ist, wenn sie einem rationalen Plan folgt. Dieser Plan wird vor den eigentlichen Handlungen festgelegt und ist dann im weiteren Verlauf stets einzuhalten. Ein Abweichen von einem bereits im Vorfeld getroffenen Plan ist laut Gauthier nicht rational.
Für das Newcombs Problem behauptet Gauthier, daß der richtige und damit rationale Plan darin besteht, sich darauf festzulegen, daß man nur die zweite Box nehmen wird (und natürlich auch an diesem Plan im weiteren Verlauf festzuhalten). Nur durch diesen Plan ist laut Gauthier gewährleistet, daß sich in der zweiten Box die eine Million Dollar befinden. Er geht in diesem Fall davon aus, daß der Predictor allwissend ist und damit die Vorhersagewahrscheinlichkeit bei 1 liegt, d.h. wenn unser Plan festliegt, dann wird auch der Predictor dieses mit an Sicherheit grenzender Wahrscheinlichkeit so vorhersagen und darauf sein Handeln abstimmen.
Die Annahme eines solchen allwissenden Predictors ist allerdings, wie schon oben ausgeführt, als sehr kritisch anzusehen.
Wir müssen aber nicht mit dem als allwissend betrachteten Predictor argumentieren, sondern können bereits die Entscheidung sich für einen Plan zu entscheiden, der vorsieht nur die zweite Box zu nehmen, und den Entschluß an diesem Plan festzuhalten, als nicht rational ablehnen.
Würden man sich dafür entscheiden nur die zweite Box nehmen zu wollen, dann würde dies, laut Gauthier, zur Folge haben, daß mit an Sicherheit grenzender

Wahrscheinlichkeit eine Million Dollar in der zweiten Box liegen. Wenn man sich dann, mit diesem Wissen, endgültig für eine oder beide Boxen entscheiden muß, dann würde die Entscheidung am Plan festzuhalten (und nur die zweite Box zu nehmen) nicht rational sein, da man statt 1 Million und 1000 Dollar sich für nur 1 Million Dollar entscheiden würde. Auch wenn der Nutzenunterschied zwischen 1 Million und 1000 Dollar und 1 Million Dollar relativ gering ausfallen wird, so ist es wohl unbestritten, daß der größere Betrag einen größeren Nutzen stiften wird.

Berücksichtigt man dann noch, daß die Entscheidung des Predictors von seiner Vorhersagewahrscheinlichkeit abhängt, dann wäre es auch möglich, daß die zweite Box eventuell leer ist. Entscheidet man sich dann nur für die zweite (leere) Box, so hat man keinen anstatt 1000 Dollar, wenn man sich für beide Boxen entschieden hätte. Auch in diesem Fall ist also der Nutzen den beide Boxen stiften größer, als der Nutzen von Box Nummer zwei allein.

Man kann also erkennen, daß es sinnvoller ist sich für beide Boxen zu entscheiden, wenn man den Plan getroffen hat, sich nur für die zweite Box zu entscheiden, da man hier immer einen Vorteil für sich hätte, vor allem, wenn die zweite Box leer sein sollte. Es ist also nicht rational in diesem Fall an seinem Plan festzuhalten.

Wenn man sich vorstellt, daß man mit dem Plan nur die zweite Box zu nehmen, den Raum betritt und man auf alle Fälle an diesem Plan festhalten wird, weil es ja, laut Gauthier, rational ist, an seinem Plan festzuhalten:

Wenn man hier Gauthier folgen würde, müßte man selbst bei zwei durchsichtigen Boxen (man würde also sehen, ob sich die Milion Dollar in der zweiten Box befindet) seinem Plan treu bleiben. Doch welcher rationale Mensch würde hier gegen seinen offensichtlichen Nutzen entscheiden und nur die 1 Million nehmen, wenn er 1 Million und 1000 Dollar haben könnte.

2.4.1 Parfits Problem als Analogie zum Newcombs Problem

Als anderes Beispiel, daß Gauthier mit seiner Lösung, sich für einen Plan zu entscheiden und diesen beizubehalten, nicht rational ist, läßt sich das Parfits Problem (Parfit, 1984) heranziehen. Genauso wird es auch bei Barnes (1997) verwendet.

Das Parfits Problem kann hier genutzt werden, da es fast analog zum Newcombs Problem arbeitet.

Parfit beschreibt in seinem Problem einen Autofahrer (A), der nach einer Autopanne auf einer wenig befahrenen Straße in der Wildnis festsitzt. Dieser Autofahrer hat die Eigenschaft, daß er immer rational und zu seinem persönlichen Nutzen handelt.

Ein zweiter Autofahrer (B) kommt hinzu, um dem ersten Autofahrer zu helfen. Dieser Fahrer B hat die Eigenschaft, daß er Menschen genau beurteilen kann und erkennt, ob diese ihn anlügen. (Besitzt also wie der Predictor im Newcombs Problem die Fähigkeit der Allwissenheit.)

Dieser Autofahrer macht nun dem Liegengebliebenen das Angebot, daß er ihn für einen bestimmte Geldbetrag nach Hause bringen würde. Der Betrag wird erst gezahlt sobald er zu Hause ist.

Fahrer A muß nun für sich entscheiden, ob es für ihn wichtiger ist nach Hause zu kommen oder ob er in der Wildnis bleiben will, damit er nicht bezahlen muß und somit sein Geld behält. Nach dieser Entscheidung richtet sich dann die Entscheidung, ob er das Angebot des anderen Fahrers annimmt oder lieber darauf verzichtet. (Dies entspricht beim Newcombs Problem dem ersten betreten des Raumes, nur das hier eine erste Entscheidung den Predictor erst in Gange setzt, nämlich ob er es annimmt oder nicht.)

Der zweite Fahrer muß darauf die Entscheidung treffen ob der andere, wenn er das Angebot annimmt, die Wahrheit sagt und er wirklich das Geld erhält, oder ob er ihn anlügt und er nichts dafür bekommt, daß er ihn nach Hause bringt. (Bei Newcomb würde dies der Entscheidung des Predictor entsprechen.)

Sollte es dazu gekommen sein, daß sie nun beim ersten Fahrer zuhause sind,

muß sich dieser noch entscheiden ober er sein Versprechen auch einhält oder nicht. (Es ist also ein Newcombs Problem mit zwei durchsichtigen Boxen.)
Zur Verdeutlichung des Problems kann man es als Matrix ausdrücken. (siehe Tabelle 2.3).

Tabelle 2.3: Parfits Problem

	glaubt dem Versprechen nicht	glaubt dem Versprechen
"wahres Versprechen"	in der Wildnis, behält sein Geld	zu Hause, muß bezahlen
"falsches Versprechen"	in der Wildnis, behält sein Geld[2]	zu Hause, behält sein Geld

Geht man nun an das Problem auf Gauthiers Weise heran, sollte der erste Fahrer sich den Plan machen das Angebot anzunehmen. Bei diesem Plan müßte er dann aber auch bleiben. Er wäre dann gezwungen den anderen Fahrer zu bezahlen.
Bei dieser Enscheidung würde der Fahrer aber gegen seinen Nutzen entscheiden, da dieser beim Geld liegt.
Es ist also richtig sich diesen Plan zuzulegen, da man so die Möglichkeit hat sich einen ersten Nutzen zu verschaffen, in diesem Fall, daß man es schafft aus der Wildnis zu entkommen. Es ist dann aber nicht rational diesem Plan treu zu bleiben, da diese Entscheidung dann gegen den eigenen Nutzen wäre.

[2] Unterschied zum Newcombs Problem, der sich aber nicht auf die Argumentation auswirkt.

Kapitel 3

Zusammenfassung

Zusammenfassend haben wir nun erkannt, daß es verschiedene Ansätze für die Lösung des Newcombs Problem in der Entscheidungstheorie gibt.
Wenn man dem traditionellen Ansatz folgt, führt dieser zur "1 Boxer" Lösung des Problems. Diese Lösung ist allerdings nur solange rational, wie wir von einem allwissenden Predictor ausgehen können. Da dieser Predictor so aber nicht existiert, können wir auch nicht von einem solchen ausgehen und erhalten über den traditionellen Ansatz die "2 Boxer" Lösung.
Zu dieser Lösung kommen wir auch, wenn wir uns vergegenwärtigen, daß unsere endgültige Handlung, nur die zweite oder beide Boxen zu nehmen, keinerlei Einfluß auf die Entscheidung des Predictor hat, da diese ja bereits getroffen wurde, bevor wir uns endgültig entscheiden müssen. Daher ist es auch hier rational sich für beide Boxen zu entscheiden.
Weiterhin haben wir erkannt, daß Gauthier mit seiner Lösung, dem rationalen Plan zu keiner endgültig rationalen Handlung kommt. Der Plan, sich nur für die zweite Box zu entscheiden, ist zwar rational, das Festhalten an diesem und die endgültige Endscheidung nur für die zweite Box stellt dann aber keine rationale Handlung mehr dar, da diese Handlung gegen den eigenen Nutzen wäre.
Somit ist abschließend festzustellen, daß eine rationale Lösung des Newcombs Problem darin besteht, sich den Plan zu machen nur die zweite Box zu nehmen, sich dann aber endgültig doch für beide Boxen zu entscheiden, da man dadurch rational handelt und dies einem den besten Nutzen einbringt.

Literaturverzeichnis

[1] Barnes, R. Eric (1997): 'Rationality, Dispositions, and the Newcomb Paradox', Philosophical-Studies 88(1), Seite: 1-28

[2] Barnes, R. Eric (1997): 'Constraint Games and the Orthodox Theory of Rationality', Utilitas 9(3), Seite:329-349

[3] Bostrom, Nick (2001): 'The Meta-Newcomb Problem', Analysis 61(4), Seite: 309-310

[4] Gauthier, David (1989): 'IN THE NEIGHBOURHOOD OF THE NEWCOMB-RPREDICTOR', Proceedings-of-the-Aristotelian-Society 89, Seite: 179-194

[5] Hurley, S.L. (1994): 'A New Take from Nozick on Newcombs Problem and Prisoners Dilemma', Analysis 54(2), Seite: 65-72

[6] Nozick, Robert (1969): 'Newcombs Problem and Two Principles of Choice', in Nicholas Rescher (ed.), Essay in Honor of Carl G. Hempel, Reidel, Dordrecht, Holland, Seite: 114-146

[7] Parfit, Derek (1984): Reasons and Persons, Clarendon Press, Oxfor, Seite 5-12