Transcription of RC 89a-14-01

Value concepts und rational agent. Zuerst geschrieben als Ergänzung zu meiner Erwiderung auf Kaplan im Schilpp Band. Aber das würde zu lang werden. Daher lieber als Grundlage **für späteren Aufsatz!**

21. 2. 58

Wertbegriffe

I. Relativ zu einem Wertsystem

V sei eine Wertfunktion. (Das setzt nicht voraus, daß es eine Person gibt, deren Wertfunktion V ist.) Dies bedeutet, daß für jede mögliche Weltgeschichte W, V(W) eine reelle Zahl ist. Da in den folgenden Definitionen es stets nur auf Differenzen zwischen Werten von V ankommt, so können wir zwei Wertfunktionen V und V′, die sich nur durch einen festen [Betrag?] unterscheiden (für jedes W, V′(W) = V(W) + A mit konstantem A), als äquivalent ansehen.

Die Proposition q beziehe sich nur auf eine beschränkte Zeitspanne t_q und ein[e] beschränkt[e] Raumregion R_q. Dann ist V(q) wie folgt zu verstehen, wobei W_T die wahre Weltgeschichte ist:

(α) (a) Wenn q tatsächlich besteht, so ist $V(q) = V(W_T)$.
(b) Wenn q falsch ist, so $V(q) = V(W_q)$, wo W_q diejenige mögliche Weltgeschichte ist, die eintreten würde, wenn q stets [stände?].

In (b) wird eine counterfactual conditional gebraucht. Die Explikation von solchen ist noch umstritten. Für unsere Zwecke mögen die folgenden Andeutungen genügen, die aber noch verfeinert werden müßten. Wir wollen in der gegenwärtigen Verwendung nur solche counterfactuals anwenden, [crossed out words] bei denen die Bedingung q in der angegebenen Weise beschränkt ist und ferner q verträglich ist mit der Gesamtheit PL der bestehenden physikalischen Gesetze (im Sinne von § , also nicht im Sinne der von den wissenschaftlern heute

anerkannten Gesetze). Dann ist W_q diejenige mögliche Weltgeschichte, die die folgenden Bedingungen erfüllt:

(β) (a) W_q übereinstimmt mit W_T im gesamten Verlauf bevor der Zeit t_q, (b) [ebenso] zur Zeit t_q außerhalb des Raumgebietes R_q, (c) innerhalb des Raum- und Zeitgebietes t_q, R_q W_q übereinstimmt so weit wie möglich mit W_T und abweicht von W_T nur soweit wie es nötig ist um q wahr zu machen;
(d) nach der Zeit t_q stimmt W_q überein mit W_T in allen Raum-Zeit Gebieten die nicht vom vorherigen q kausal affektiert sind, während sie für die durch q beeinflußten Gebiete soweit von W_T abweicht wie es durch q zusammen mit den Gesetzen PL bestimmt ist.

(γ) p ist besser als q mit Bezug auf die Wertfunktion $V =_{Df} V(p) - V(q) > 0$.

(δ) p ist gut mit Bezug auf die Wertfunktion $V =_{Df}$ p ist besser als nicht-p.

Angenommen, ein Agent X hat zur Zeit t die Wahl zwischen den möglichen Aktionen einer Menge A_X. Wir definieren:

(ε) Die möglichen Aktion a von A_X ist ein Optimum inbezug auf die Wertfunktion V $=_{Df}$ keine Aktion von A_X ist besser (im Sinne von (γ)) als a mit Bezug auf V.

(22. 2.)

Eine Person X hat zu einer gegebenen Zeit nicht nur eine Wertfunktion, sondern eine ganze Reihe von solchen, die verschiedene Wertaspekte repräsentieren. Wenn X, eingedenk der [Diät-] Ratschläge seines Doktors, sagt: "Es ist besser für mich, eine Speise zu vermeiden", so hat er im Sinne eine gewisse Wertfunktion, die nur die Gesundheitswerte und zwar für sich Selbst darstellt. Andre teilweise Wertaspekte mögen sein: sein geschäftlicher Profit, sein esthetisch[er] pleasure, sein eigenes Wohlergehen inbezug auf alle Hinsichten zusammen, das

Wohlergehen einer Familie, das einer großen Gruppe, das der Nation, das der Menschheit. Da aber gibt es eine allumfassende Wertfunktion des X, die alle Aspekte umfaßt, und in der dann auch zum Ausdruck gebracht ist, welches relative Gewicht der eine und der andere Aspekt in irgendeiner möglichen Gesamtsituation haben soll, Aspekte die zuweilen in Konflikt miteinander sind. Unter einer "moralischen Wertung" wird verschiedene Dinge verstanden. Vielleicht ist es am besten, diesen Term zu gebrauchen für die Gesamtbewertung, in der die verschiedenen Aspekte mit enthalten sind.

II. Der rationale Agent.

(ζ) Relative Rationalität. Inbezug auf eine Wertfunktion V, eine Cred-Funktion Cred, eine Evidenz E und eine Menge A von möglichen Aktionen ist eine Aktion a von A rational $=_{Df}$ für keine von a verschiedene Aktion a′ von A ist die Vorziehung von V(W) mit Hilfe von Cred auf der Basis von E und a′ größer als auf der Basis E und a. (Die Vorziehung von V(W) inbezug auf eine gewisse Evidenz ist die Summe über alle möglichen W der Produkte von V(W) mit der cred. von W auf der Basis der Evidenz in Frage, siehe § . . .)

Es gibt gewisse Standards durch die eine Cred-Funktion als unvernünftig kritisiert werden kann; diese sind an anderer Stelle besprochen worden (Kemeny's essay §[III]; und meine §[26(IV)] in dieser Reply). Diese Standards aufzustellen ist die Aufgabe der induktiven Logik.

Gibt es auch standards von Rationalität für Wert-Funktionen? Die obengenannten Standards der induktiven Logik sind hier nicht anwendbar. Die Akzeptierung einer Wertfunktion ist gänzlich unabhängig von faktischen Fragen, weil die Wertfunktion primär bewertet nicht einzelne Handlungen oder Vorgänge sondern vielmehr ganze möglichen Weltgeschichten. Überlegungen über zu erwartende Konsequenzen einer Handlung kommen hier gar nicht ins [Spiel], denn in einer W sind alle Folgen schon mitgegeben und angenommen, die Funktion V_1 bewertet W_1 höher als W_2, während die Funktion V_2 das Umgekehrte tut:

(a) $V_1(W_1) > V_1(W_2)$

(b) $V_2(W_1) < V_2(W_2)$.

[?Angenommen??], der Agent X_1 akzeptiert V_1 und X_2 akzeptiert V_2. Angenommen, X_1 und X_2 diskutieren zusammen über ihre Wertfunktionen und, in particular, über die deskriptiven Ergebnisse (a) und (b). Bei ihrer Diskussion betrachten sie also nur die beiden Geschichte W_1 und W_2. X_1 mag für each von diesen beiden Geschichten andere Evidenzwerte haben als X_2; aber das ist irrelevant für die Frage der Wahl zwischen V_1 und V_2. Diese Frage ist nur, ob man W_1 für besser bewertet als W_2 oder vielmehr als schlechter; das had nichts zu tun mit der Frage, ob W_1 eintreffen wird oder größere Wahrscheinlichkeit hat als W_2.

Obwohl also Logik, einschließlich induktiver Logik, und faktische Erkenntnisse irrelevant sind, scheint es mir doch, das es noch andere, reine valuational, Gründe gibt, nach denen man eine Wertfunktion beurteilen kann als mehr rational oder weniger rational als eine andere. Ich will nicht versuchen, hier fundamentale Prinzipien für eine solche Beurteilung aufzustellen. Ich will nur einige Gesichtspunkte erwähnen, deren Berechtigung bei einer solchen Beurteilung plausibel erscheint und wohl von den meisten gebilligt werden würden, auch wenn sie in ihren Wertungen voneinander stark abweichen. Erstens scheint es vernünftig, zu fordern, daß eine Wertfunktion V(W) ableitbar ist aus allgemeinen Prinzipien über die Bewertung von Einzelvorgängen; und zwar so, daß der Wert von V(W) eine algebraische Summe (oder Integral) ist von positiven oder negativen Werten, die nach irgendwelchen Prinzipien für gewisse sehr spezielle Vorgänge bestimmt werden, während die übrigen Vorgänge irrelevant sind. (Die relevanten Vorgänge kommen z.B. aus gewissen Vorgängen von Gefühlen in Menschen, oder eine allgemeinere Art von Vorgängen in beings die animate sind oder für solche angesehen werden; während die Vorgänge in der [?Anorganität] natürlich irrelevant sind.) Ferner wird zu fordern sein, daß die Prinzipien einen generellen Charakter haben; ausdrückbar sind durch mathematische Funktionen der relevanten Züge der betreffenden Vorgänge, und zwar mathematische Funktionen, die stetig sind und verhältnismäßig glatt, vielmehr als jumping auf und nieder. Diese Beispiele von Forderungen mögen zweifelhaft sein. Ich habe sie nicht erwähnt, um ihre Gültigkeit zu behaupten, sondern nur, um anzudeuten warum ich glaube, daß es gewisse Standards gibt, die eine Wertfunktion erfüllen muß um rational zu sein. Die Klarstellung solcher Standards kann ich hier nicht versuchen. Aber es scheint klar, daß, wenn solche Standards ausgearbeitet werden würden, sie nur gewisse Wertfunktionen als irrational

ausschalten würden, aber doch noch eine unendliche Menge von verschiedenen Wertfunktionen zulassen würden, die außerordentlich verschieden voneinander sind, und darunter viele, die von den meisten Menschen und vielleicht von allen, als gänzlich verkehrt und immoral[isch] angesehen werden würden. Die Standards, von denen ich gesprochen habe, haben also keineswegs die Funktion, "Immoralität" auszuschalten oder eine Unterscheidung zwischen den Werturteilen, die psychologisch in Controversen über Moral oder politische Fragen vorkommen, zu treffen. Im folgenden werde ich sprechen von "den Standards der Rationalität für Wertfunktionen", als ob sie schon aufgestellt wären.

Nun wollen wir definieren:

Das Verhalten eines Agenten X is perfekt rational während einer gewissen Zeitperiode Δt, wenn er folgende Bedingungen erfüllt:

(η) (a) Im deduktiven Denken, was die ganze reine Mathematik einschließt, macht er während Δt niemals Fehler.

(b) Während der Zeitperiode Δt verwendet er in seinem induktiven Denken eine rationale Methode; mehr spezifisch, es gibt für ihn während dieser Periode eine cred[i]b[ilit]y Funktion Cred_X, die die Bedingungen der Rationalität erfüllt.

(c) Sein Verhalten während der Periode Δt ist bestimmt (in der Weise zu beschreiben unter (d)) durch eine Wertfunktion V_X, die alle Standards von Rationalität erfüllt.

(d) Wenn immer X zu eine[m] Zeitpunkt t innerhalb der Periode d_X die Wahl hat zwischen verschiedenen Aktionen einer Menge $A_{X,t}$, und wenn zu t seine gesamte Evidenz $E_{X,t}$ ist, so hat die von X beschlossene Aktion a relative Rationalität (im Sinne von ζ) inbezug auf V_X, Cred_X, $E_{X,t}$, und $A_{X,t}$.

[An]genommen X ist perfekt rational zur Zeit t und wählt die Aktion a von A_X. Dann ist es trotzdem möglich, daß a nicht ein Optimum inbezug auf V_X ist. Es mag sein, daß eine Aktion a′ besser ist als a inbezug auf V_X, due zu gewissen Umständen, die dem X zur Zeit der Handlung nicht bekannt sind. Es mag sogar sein, daß die objektiv bessere, d.h. erfolgreichere Aktion a′ nicht rational für X wäre. Wie anderswo betont (§ . . .), ist Rationalität nicht zu bestimmen durch den Erfolg.

Kein Mensch ist jemals perfekt rational in dem soeben definierten Sinn. "Mehr rational", sei es angewendet auf verschiedene Perioden, oder auf zwei mögliche Verhalten derselben Personen in derselben Periode, ist wohl kaum exakt definierbar. Roh gesprochen, ist ein Verhalten mehr rational als ein anderes, wenn es dem perfekt rationalen Verhalten näherkommt. Aber da Abweichungen vom perfekt rationalen Verhalten in ganz verschiedenen Hinsichten möglich ist, z.B. in den oben (η) genannten Hinsichten (a), (b), (c), und (d) und innerhalb jeder von diesen wiederum in verschiedenen Hinsichten, so ist es wohl kaum ohne willkürliche Festsetzung möglich zu bestimmen, unter welchen Bedingungen eine Abweichung in einer gewissen Hinsicht als gleich gelten soll zu einer Abweichung in einer gewissen anderen Hinsicht.