Epistemische Logik

und die Frage, ob man wissen kann, dass man nicht weiß

Manuel Bremer

Einführende Bemerkungen

1. Mein Thema liegt an der Schnittstelle zwischen Erkenntnistheorie und philosophischer Logik (hier genauer: der modalen epistemischen Logik).

2. Die epistemische Logik, die hier betrachtet wird, ist die klassische epistemische Logik, die analog zur alethetischen Modallogik epistemische Operatoren (wie „W“ für Wissen, „M“ für Meinen, „Ü“ für Überzeugtsein) einführt. Die epistemische Logik im Allgemeinen hat sich in eine Fülle von Ansätzen weiterentwickelt (z.B. nicht-monotone epistemische Logiken). Der klassische Ansatz ist aber immer noch lebendig.

3. Es werden zwei Unmöglichkeitstheoreme im strengen Sinne bewiesen und eine Inkompatibilität zwischen einigen der untersuchten Begriffe wird plausibilisiert.

Positive und Negative Introspektion in der Epistemischen Logik

- Wenn ein Subjekt die Meinung hat “Die Katze ist auf der Matte”, ist die introspektive Meinung “Ich meine, dass die Katze auf der Matte ist”.

- Eine Frage, die sich nun stellt, ist: Wieviel (verlässlichen) Zugang haben epistemische Subjekte zu ihren epistemischen Zuständen? [Zuständen ‚erster Ordnung’ (ohne epistemische Operatoren)]

- Im Folgenden geht es insbesondere um die sogenannte „Negative Introspektion“, die zusammenhängt mit der Frage, ob man immer weiß, dass man nicht weiß.

Wenn „M“ Meinen ausdrückt, dann bestünde ideale Selbstwahrnehmung in den folgenden beiden Prinzipien (für irgendwelche Sätze/Behauptungen α):

(i) *Positive Introspektion*: $M\alpha \supset MM\alpha$

(ii) *Negative Introspektion*: $\neg M\alpha \supset M\neg M\alpha$

Darüber hinaus werden in der epistemischen Logik zwei Idealisierungen zum logischen Abschluss von Meinungen (und anderen epistemischen Zuständen) gemacht:

(iii) $\vdash \alpha \Rightarrow M\alpha$

(iv) $\vdash (\alpha \supset \gamma), M\alpha \Rightarrow M\gamma$

Ein paar kurze Bemerkungen zu den Introspektionsprinzipien

• Diese Bedingungen korrespondieren zu den üblichen modallogischen Prinzipien. Die beiden Introspektionsprinzipien korrespondieren zu den charakteristischen Axiomen der Systeme S4 und S5. [$\Box\alpha \supset \Box\Box\alpha$ bzw. $\Diamond\alpha \supset \Box\Diamond\alpha$ ($\equiv \neg\Box\neg\alpha\supset\Box\neg\Box\neg\alpha$)]

• Alle diese Bedingungen (bzw. entsprechende Axiome) sind umstritten, insofern sie erkenntnismäßig zu viel forderten oder psychologisch unrealistisch seien. An diesen Kritiken ist vieles berechtigt, hier soll es aber um noch grundsätzlichere Probleme mit der Negativen Introspektion gehen.

• Für endliche Datenbanken, die einzelne Fakten speichern, scheint Positive Introspektion – evtl. sogar Negative Introspektion – nicht problematisch. Hier interessiert jedoch der allgemeinere Fall.

Negative Introspektion kann nicht algorithmisch sein!

- Vorklärung:

Die Church-Turing-These: *Alles was intuitiv berechenbar ist, ist Turing-berechenbar.*

Informeller: Alle Prozeduren, die (i) aus kleinen („geistlosen“) Schritten bestehen und (ii) im Falle, dass sie ein Ergebnis (ausgehend von einer endlichen Eingabe) liefern, dies in endlicher Zeit liefern und (iii) substratneutral sind, sind „algorithmische“ Prozeduren. Diese Prozeduren kann eine Turingmaschine ausführen/berechnen.

Kognitionswissenschaftliche Lesart: Alle kognitiven Fähigkeiten (wie das Verstehen von Wörtern, das Klassifizieren von Objekten…) sind *auf irgendeiner Beschreibungsebene* (prozedural oder neurocomputational…) algorithmisch.

- Nicht-algorithmischen Fähigkeiten haftet der Nymbus des Mysteriösen an.

1. Resultat: Negative Introspektion kann nicht algorithmisch sein.

Beweis: Angenommen wir hätten ein kognitives System A, das über keine kontingenten Meinungen verfügt, sondern nur Meinungen über logischen Abschluss gewinnt (Bedingungen (iii) und (iv)) sowie über Introspektion (Bedingungen (i),(ii)) verfügt. Sei Δ irgendeine *unentscheidbare* Logik, deren Theorem aufzählbar sind (abgeleitet werden können), etwa die Prädikatenlogik Erster Stufe. A kann Δ entscheiden! $\lightning$

<u>Prozedur:</u>

Die Sätze der Sprache von Δ sind aufzählbar (von einer TM M_1). Wir lassen M_1 A einen Satz α vorlegen. A überprüfe: $M\alpha$? Entweder wird der Satz für wahr gehalten oder nicht [Meinungen hat man oder hat man nicht: $M\alpha \vee \neg M\alpha$, nicht zu verwechseln mit: $M\alpha \vee M\neg\alpha$, was nicht gilt].

Wenn der Satz von A geglaubt wird, weiß A das aufgrund positiver Introspektion. Da Meinungen, nach Definition von A, nur durch logischen Abschluss gewonnen werden, weiß A nun, dass $\vdash\alpha$. Wird der Satz nicht geglaubt, weiß A dies aufgrund negativer Introspektion. Da logischer Abschluss der einzige Weg zu Meinungen für A ist, weiß A, dass dieser Weg nicht beschritten wurde, also per Kontraposition $\nvdash\alpha$.

D.h. jeder Satz α von Δ kann von A bezüglich $\vdash$ entschieden werden! ■

Verschärfung: Die Theoreme von Δ sind aufzählbar durch eine TM M_2. Das heißt auf positive Introspektionen können wir bei A sogar verzichten! Alles liegt an Negativer Introspektion■

Erläuterung:

Das geschilderte Vorgehen liefert *kein* Entscheidungsverfahren im strikten Sinne (und widerpricht damit nicht den limitativen Theoremen [wie *Churchs Theorem*]). Anders herum: Es *kann sich* wegen dieser limitativen Theoreme nicht um ein Entscheidungsverfahren handeln. Das heißt für die „kritische Zutat“ Negative Introspektion: Negative Introspektion *kann nicht* algorithmisch sein! (Sonst hätten wir ein solches Verfahren.)

Negative Introspektion und Wissen als wahre Überzeugung

Introspektionsprinzipien sind *kontrovers* im Lichte der Erkenntnistheorie, sie sind *desaströs* in Kombination mit einem starken Wissensbegriff (Wissen als wahre Überzeugung [eine Überzeugung ist eine Meinung, die man für *gewiss* hält]):

(W_+) $\quad W_+\alpha \equiv \ddot{U}\alpha \wedge \alpha$

Statt:

(W_F) $\quad W_F\alpha \equiv \ddot{U}\alpha \wedge \alpha \wedge F\alpha$

Die Logik des Wissens und die des Überzeugseins werden traditionell als „normale“ Modallogiken eingeführt:

(T) $W_+\alpha \supset \alpha$ [natürlich nicht für „Ü“: $\nvdash Ü\alpha \supset \alpha$]

(K) $W_+(\alpha \supset \gamma) \supset (W_+\alpha \supset W_+\gamma)$ [$Ü(\alpha \supset \gamma) \supset (Ü\alpha \supset Ü\gamma)$] (KÜ)

(RW_+) $\vdash \alpha \Rightarrow \vdash W_+\alpha$ [$\vdash \alpha \Rightarrow \vdash Ü\alpha$] (RÜ)

Haben wir positive Introspektion für Ü ((Ü4), entsprechend Bedingung (i)) erhalten wir positive Introspektion für starkes Wissen.[1]

(S4) $W_+\alpha \supset W_+W_+\alpha$

[1] “$W_+\alpha$“ ist “$Ü\alpha \wedge \alpha$“ das impliziert durch (Ü4) “$ÜÜ\alpha$“ und sich selbst. Mit dem (K) Axiom für “Ü” (von rechts nach links angewandt $[(KÜ) \equiv (Ü(\alpha \wedge \gamma) \equiv Ü\alpha \wedge Ü\gamma)]$) erhalten wir: $Ü(Ü\alpha \wedge \alpha) \wedge (Ü\alpha \wedge \alpha)$, d.h. $W_+W_+\alpha$.

Wie sieht es aus mit Negativer Introspektion für starkes Wissen?

Dies wäre:

(S5*) $\neg W_+\alpha \supset W_+\neg W_+\alpha$

Bzw.:

(S5*') $\neg (\ddot{U}\alpha \wedge \alpha) \supset \neg(\ddot{U}\alpha \wedge \alpha) \wedge \ddot{U}(\neg(\ddot{U}\alpha \wedge \alpha))$

2. Resultat: PC∪{(W_+),(Ü4),(KÜ), (S5*)} ist inkonsistent!

Beweis: Angenommen wir wissen α nicht, weil ¬α, sind aber überzeugt: Üα**.** Das Vorderglied von (S5*') $[\neg(Ü\alpha \wedge \alpha) \supset \neg(Ü\alpha \wedge \alpha) \wedge Ü(\neg(Ü\alpha \wedge \alpha))]$ ist dann wahr, also haben wir das Hinterglied. Wir haben also das zweite Konjunkt: Ü¬(Üα∧α).

Aussagenlogisch ist $\neg(Ü\alpha \wedge \alpha)$ äquivalent zu $\alpha \supset \neg Ü\alpha$.

Ersetzen wir diese Äquivalente erhalten wir $Ü(\alpha \supset \neg Ü\alpha)$.

Aufgrund von (KÜ): (2) $Ü(\alpha \supset \neg Üa) \supset (Ü\alpha \supset Ü\neg Ü\alpha)$

Das heißt: in unserem Fall können wir, da das Vorderglied vorliegt und außerdem Üα, zweimal abtrennen, und erhalten, indem wir wieder Üα als Konjunkt hinzunehmen:

(3) $Ü\alpha \wedge Ü\neg Ü\alpha$

In Kombination mit (Ü4) erhalten wir schließlich einen Widerspruch:

(4) $ÜÜ\alpha \wedge Ü\neg Ü\alpha$ oder, wieder mit (KÜ): $Ü(Ü\alpha \wedge \neg Ü\alpha)$

Also muss (S5*) zurückgewiesen werden. Nicht allein, weil wir eine falsche Überzeugung haben, haben wir eine widersprüchliche Überzeugung $C\bot$. ■

Weiterführung

Lenzen (1978, 1979, 1980) ist nicht zufrieden mit **S4** als Logik von W_+. Er betrachtet eine abgeschwächte Form der Negativen Introspektion, welche dem charakteristischen Axiom der Modallogik **S4.4**. entspricht.

(S4.4) $\alpha \supset (\neg W_+ \neg W_+ \alpha \supset W_+ \alpha)$

Bzw.:

(S4.4') $\alpha \supset (\neg W_+ \alpha \supset W_+ \neg W_+ \alpha)$

Die zweite Version macht deutlich, dass es sich um eine abgeschwächte Variante von allgemeiner Negativer Introspektion handelt: (S5**) $\alpha \vee \neg \alpha \supset (\neg W_+ \alpha \supset W_+ \neg W_+ \alpha)$

3. Resultat: PC $\cup$ {(W_+),(Ü4),(T),(K),(KÜ),(S4.4)} ist unakzeptabel.

Die Logik von W_+ muss schwächer sein als S4.4

Argument: $W_+\alpha$ impliziert $\alpha \wedge \neg W_+\neg W_+\alpha$.[2] Mit (S4.4) gibt das:

(5) $W_+\alpha \equiv \alpha \wedge \neg W_+\neg W_+\alpha$

Schreibt man die Definition des starken Wissensbegriff darunter

(W_+) $W_+\alpha \equiv \alpha \wedge \dot{U}\alpha$

ist klar, dass wir folgendes Theorem haben:

(6) $\dot{U}\alpha \equiv \neg W_+\neg W_+\alpha$

[2] Da $W_+\alpha$ durch (S4) $W_+W_+\alpha$ impliziert, gäbe $W_+\neg W_+\alpha$ einen Widerspruch aufgrund des (K) Axioms.

(6) $\ddot{U}\alpha \equiv \neg W_+ \neg W_+ \alpha$

ist bizarr: Angenommen wir haben noch nie über α nachgedacht (oder α ist jenseits menschlichen Erkennens). Dann wissen wir auch nicht α: $\neg W_+ \alpha$. Und da wir noch nie über α nachgedacht haben, haben wir auch keine introspektiven Fragen an uns gestellt: $\neg W_+ \neg W_+ \alpha$. Aber jetzt sagt uns (6), dass wir dann überzeugt sind, α sei wahr! Für einen beliebigen Sachverhalt jenseits unseres Erkennens verpflichtet uns S4.4 für W_+ zu einer Überzeugung! Das ist (erkenntnistheoretisch) völlig inakzeptabel.

Überblick über weitere Konsequenzen

- Der starke Wissensbegriff erwies sich als inkompatibel mit bedingten oder unbedingten Prinzipien Negativer Introspektion. Wie sieht es mit Negativer Introspektion *beim Überzeugungsbegriff* aus?

- Wenn man die Definition (W_+) verwendet, lässt sich die Logik des Überzeugtseins als *äquivalent* zur oben zurückgewiesenen Logik S4.4 für W_+ erweisen! Der entscheidende Schritt im Äquivalenzbeweis beruht auf (Ü5) [$\neg Ü\alpha \supset Ü\neg Ü\alpha$], d.h. auf Negativer Introspektion für Überzeugtsein. (Ü5) muss also auch zurückgewiesen werden.

Zusammenfassung und Schlussbemerkungen

Negative Introspektion ist nicht nur eine sehr starke Annahme für menschliche Räsonierer.

1. Negative Introspektion *kann kein algorithmischer Vorgang sein* (gegeben die Möglichkeit von logischer Abgeschlossenheit von Meinungen bezüglich einer Logik).

2. Allgemeine Negative Introspektion *ist inkonsistent* im Rahmen einer normalen Modallogik für den starken Wissensbegriff.

3. Abgeschwächte Negative Introspektion *führt* im Rahmen einer normalen Modallogik für den starken Wissensbegriff *zu absurden Ergebnissen*.

   [Bemerkenswerterweise werden trotzdem entsprechende Logiken vorgeschlagen.]

Erkenntnistheoretisch kann man vielleicht Folgendes sagen:

1. Es zeigt sich einmal mehr, dass „negative“ Fähigkeiten komplexer sind als „positive“ Fähigkeiten. Positive Introspektion führt nicht zu entsprechenden Ergebnissen.
2. Die Komplexität Negativer Introspektion ähnelt dem Komplexitätsunterschied zwischen Beweisbarkeit ($\sum_1$-Satz) und Nichtbeweisbarkeit ($\prod_1$-Satz).
3. Die Inkompatibilität lässt sich nicht so einfach zeigen für den Begriff fundierten Wissens, der dadurch wieder mehr an Glaubwürdigkeit gewinnt.
4. Sokrates kann manchmal wissen, dass er nicht weiß – aber nicht immer!

Verweise

Lenzen, Wolfgang (1978). *Recent Work in Epistemic Logic*. Special Issue of Acta Philosophica Fennica, Vol. 30 (1).

- (1979). "Epistemologische Betrachtungen zu [S4, S5]", *Erkenntnis*, 14, pp.33-56.

- (1980). *Glauben, Wissen und Wahrscheinlichkeit*. System der epistemischen Logik. Wien/New York.

Meyer, J.-J. Ch./van der Hoek, W. (2004). *Epistemic Logic for AI and Computer Science*. Cambridge.