Penultimate draft of an article published in: Hoffmann-Kolss, V., Rathgeb, N. (eds), *Handbuch Philosophie des Geistes*. Stuttgart: J.B. Metzler, 2024, pp. 425-434. https://doi.org/10.1007/978-3-476-05416-6_39

Selbstwissen und die Autorität der ersten Person

1. Grundsätzliches

Unter ›Selbstwissen‹ wird in der zeitgenössischen Diskussion das Wissen verstanden, das eine Person über ihre eigenen gegenwärtigen mentalen Zustände besitzt. Da es eine Vielzahl unterschiedlicher Typen mentaler Zustände gibt, variiert der Gehalt der Überzeugungen, in denen sich Selbstwissen manifestiert, erheblich: Überzeugungen mit dem Gehalt ›Ich habe Schmerzen‹ gelten ebenso gut als Kandidatinnen für Selbstwissen wie Überzeugungen, deren Gehalte mit Hilfe der Sätze ›Ich glaube, dass Deutschland in Europa liegt‹, ›Ich habe den visuellen Eindruck von etwas Rotem‹ oder ›Ich möchte ein Eis essen‹ zum Ausdruck gebracht werden. Der kleinste gemeinsame Nenner von Überzeugungen, die als Kandidatinnen für Selbstwissen in Frage kommen, besteht darin, dass sich ihre Gehalte als Instanzen des Schemas ›Ich befinde mich in φ‹ auffassen lassen – wobei die Wendung ›sich in φ befinden‹ als Platzhalter alltagspsychologischer Prädikate (wie ›Schmerzen haben‹, ›glauben, dass Deutschland in Europa liegt‹, ›den visuellen Eindruck von etwas Rotem haben‹, ›ein Eis essen möchten‹ usw.) dient. Im Folgenden sollen solche Überzeugungen ›Selbstüberzeugungen‹ genannt werden.

Selbstüberzeugungen sind aus zweierlei Gründen in den Fokus der philosophischen Debatte geraten: zum einen, weil sie epistemische Merkmale zu haben scheinen, die wir so bei anderen Überzeugungen nicht antreffen, und zum anderen, weil sprachliche Äußerungen der Form ›Ich befinde mich in φ‹ von ihren Adressaten auf eine charakteristische Weise bevorzugt behandelt zu werden scheinen.

2. Privilegierter Zugang und Autorität der ersten Person: Vorklärungen

Zu den besonderen epistemischen Merkmalen von Selbstüberzeugungen werden traditionellerweise ihre Direktheit und ihre besondere Sicherheit gezählt. Unter ›Direktheit‹ wird dabei der Umstand verstanden, dass Selbstüberzeugungen – anders als Überzeugungen, die von den mentalen Zuständen anderer handeln – nicht das Resultat von Schlussfolgerungen darstellen, die auf Verhaltensbeobachtungen beruhen. Und die Rede von ›Sicherheit‹ soll auf den Umstand hindeuten, dass Selbstüberzeugungen zuverlässiger und weniger anfällig für bestimmte Arten von Fehler sind als andere Überzeugungen. Ich selbst – so scheint es zumindest – kann nicht nur in einer direkten, von der sinnlichen Wahrnehmung unabhängigen Weise Wissen über meine eigenen mentalen Zustände erlangen, sondern ich weiß auch *besser* als alle anderen, was in meinem Geist vor sich geht. In der Literatur werden beide Merkmale, Direktheit und Sicherheit, gerne unter dem Stichwort ›privilegierter Zugang‹ zusammengefasst.

Für die charakteristische Weise, in der sprachliche Äußerungen der Form ›Ich befinde mich in φ‹ bevorzugt behandelt werden, hat sich dagegen die Rede von einer ›Autorität der ersten Person‹ eingebürgert. Damit ist gemeint, dass Hörerinnen Äußerungen der Form ›Ich befinde mich in φ‹ ein höheres Maß an Vertrauen schenken als Äußerungen der Form ›Er/sie befindet sich in φ‹. Angenommen, dass mein Gesprächspartner mit Blick auf eine Frau am Nebentisch zu mir sagt: ›Sie hat Kopfschmerzen und würde gerne das Restaurant verlassen‹. Es wäre nicht ungewöhnlich, wenn ich in dieser Situation zurückfragen würde: ›Wirklich? Woher weißt du das?‹ Würde jedoch mein

Gesprächspartner zu mir sagen: ›Ich habe Kopfschmerzen und würde gerne das Restaurant verlassen‹, dann klänge meine Rückfrage ›Wirklich? Woher weißt du das?‹ irgendwie unangemessen. Normalerweise schenken wir Äußerungen der Form ›Ich befinde mich in φ‹ ohne weiteres Glauben. Zweifel oder das Einfordern von Gründen scheinen unangebracht zu sein.

Ein Großteil der philosophischen Diskussion über Selbstwissen dreht sich um die Frage, wie sich der Umstand erklären lässt, dass wir privilegierten Zugang zu unseren eigenen mentalen Zuständen haben bzw. dass Äußerungen der Form ›Ich befinde mich in φ‹ eine besondere Autorität genießen. Oftmals steht jedoch auch die Frage im Mittelpunkt, wie sich die Rede von ›Direktheit‹ und ›Sicherheit‹ näher ausbuchstabieren (vgl. Alston 1971) bzw. wie sich die Idee der Autorität der ersten Person genauer fassen lasse (vgl. Barz 2018). Und manchmal wird sogar in Frage gestellt, ob wir tatsächlich privilegierten Zugang zu den eigenen mentalen Zuständen haben (vgl. Ryle 1949) bzw. ob Äußerungen der Form ›Ich befinde mich in φ‹ tatsächlich eine besondere Autorität genießen (vgl. Doppelt 1978). Wie so häufig in philosophischen Debatten ist so gut wie alles strittig: nicht nur, worin die beste Erklärung für das Phänomen, sondern auch, worin das zu erklärende Phänomen besteht – und sogar, ob das zu erklärende Phänomen überhaupt existiert.

3. Das cartesianische Bild

Um sich einen Überblick über die Debatte zu verschaffen, ist es hilfreich, mit der Betrachtung zweier traditioneller Thesen zu beginnen, die – ob zu Recht oder zu Unrecht, sei hier dahingestellt – häufig René Descartes zugeschrieben werden (s. Kap. 3):

[Unfehlbarkeit] Notwendigerweise gilt: Wenn Person *S* zum Zeitpunkt *t* glaubt, dass sie sich in φ befindet, dann befindet sich *S* zu *t* in φ.

[Phosphoreszenz] Notwendigerweise gilt: Wenn sich Person *S* zum Zeitpunkt *t* in φ befindet, dann glaubt *S* zu *t*, dass sie sich in φ befindet.

Zusammengenommen ergeben die beiden Thesen ein Bild des Geistes, das gemeinhin als ›cartesianisch‹ bezeichnet wird. Dem cartesianischen Bild zufolge besteht eine überaus enge Verbindung zwischen einem mentalen Zustand φ und der Überzeugung, dass φ vorliegt: Bereits das bloße Vorliegen von φ sei hinreichend für das Haben dieser Überzeugung. Eine epistemische Anstrengung, ein besonderer Akt der Aufmerksamkeit, ein Suchen, Abtasten, Finden, Fokussieren oder In-den-Blick-Nehmen, sei nicht vonnöten. Sobald die Überzeugung, dass man sich in φ befinde, gebildet sei, bestehe zudem eine Garantie ihrer Wahrheit: Eine wie auch immer geartete Differenz zwischen dem, was wir über das Vorliegen von φ glauben, und dem, wie es sich tatsächlich mit φ verhält, sei ausgeschlossen.

Das cartesianische Bild legt nahe, dass sich unser epistemisches Verhältnis zu unserem eigenen Geist in fundamentaler Weise von unserem epistemischen Verhältnis zur Außenwelt unterscheidet. Während wir materielle Gegenstände in der Außenwelt – falls wir überhaupt von ihnen Notiz nehmen – lediglich vermittels sinnlicher Erscheinungen erfassen können, die von der Wirklichkeit mitunter erheblich abweichen, scheint der Geist eine Sphäre vollkommener Durchsichtigkeit zu sein, in der die Differenz zwischen Erscheinung und Wirklichkeit aufgehoben ist: So wie uns unsere mentalen Zustände erscheinen, sind sie auch beschaffen – und so wie unsere mentalen Zustände beschaffen sind, erscheinen sie uns auch.

4. Kritik am cartesianischen Bild und die Grundzüge der zeitgenössischen Debatte

Das cartesianische Bild mag im Zusammenhang mit Schmerzen plausibel erscheinen. Sobald wir jedoch andere Arten mentaler Zustände, etwa Überzeugungen, Wünsche oder Emotionen, in den Blick nehmen, verliert es schnell an Überzeugungskraft. Spätestens seit den Arbeiten Sigmund Freuds wissen wir, dass wir Überzeugungen, Wünsche oder Emotionen besitzen können, die sich systematisch unserem Zugriff entziehen – und dass das Selbstbild, das wir von uns haben, häufig nicht mit der Realität übereinstimmt (s. Kap. 40). Das cartesianische Bild scheint nicht einmal im Fall von Schmerzen angemessen zu sein. Man denke nur an das berühmt-berüchtigte Beispiel des Initiationsrituals, bei dem einer Person die Augen verbunden werden und ihr gesagt wird, sie werde mit einem glühenden Schürhaken gebrandmarkt werden. Anstelle eines glühenden Schürhakens wird jedoch ein Eiswürfel auf die nackte Haut gepresst. In diesem Fall mag es der Person so erscheinen, als ob sie einen brennenden Schmerz spüre, sie mag (wenngleich nur für einen kurzen Moment) *glauben*, sie habe Schmerzen – in Wirklichkeit hat sie jedoch lediglich eine Kälteempfindung. Dieses Beispiel wirft nicht nur ein schlechtes Licht auf die Unfehlbarkeitsthese, sondern ist auch geeignet, die Phosphoreszenzthese zu entkräften. Denn die beschriebene Person hat zwar in dem Moment, in dem ihr der Eiswürfel auf die nackte Haut gepresst wird, eine Kälteempfindung, glaubt jedoch (zumindest in diesem Moment) nicht, dass sie eine Kälteempfindung hat.

Es sieht daher so aus, als sei das epistemische Verhältnis, in dem wir zu den Zuständen und Vorgängen unseres eigenen Geistes stehen, nicht so eng, wie uns das cartesianische Bild glauben machen will. Eine Möglichkeit, sich die zeitgenössische Diskussion über Selbstwissen zurechtzulegen, besteht darin, sie als Streitgespräch zwischen zwei Lagern aufzufassen: Auf der einen Seite stehen Philosophen, die bestrebt sind, nachzuweisen, dass sich das epistemische Verhältnis, in dem wir zu unseren eigenen mentalen Zuständen stehen, nicht in fundamentaler Weise vom epistemischen Verhältnis unterscheidet, das wir zu materiellen Gegenständen in der Außenwelt unterhalten. Auf der anderen Seite stehen diejenigen, die zwar einräumen, dass Ignoranz und Irrtum bezüglich der eigenen mentalen Zustände möglich sei, andererseits aber darauf pochen, dass das epistemische Verhältnis, in dem wir zu unseren eigenen mentalen Zuständen stehen, grundsätzlich anderer Art sei als das epistemische Verhältnis, in dem wir zu materiellen Gegenständen in der Außenwelt stehen.

5. Variationen der Phosphoreszenz I

Ein beispielhafter Vertreter des ersten Lagers ist David Armstrong (1968). Als Anhänger der funktionalistischen Version der Identitätstheorie (s. Kap. 9, 14) geht Armstrong davon aus, dass mentale Zustände nichts anderes als Gehirnzustände sind. Er hält Selbstwissen daher für das Produkt eines neuronal implementierten Mechanismus, mit dessen Hilfe das Gehirn seine eigenen Zustände scannt. Im Prinzip, so Armstrong, unterscheidet sich die Arbeitsweise dieses Mechanismus nicht von der sinnlichen Wahrnehmung materieller Gegenstände.

Sydney Shoemaker (1994), der als paradigmatischer Vertreter des zweiten Lagers gelten kann, hält die Analogie zwischen dem Erwerb von Selbstwissen und sinnlicher Wahrnehmung für verfehlt. Sein bekanntestes Argument besteht darin, dass, falls die Analogie tragen sollte, eine normal intelligente, begrifflich kompetente und rationale Person denkbar sein müsste, die nicht in der Lage wäre, Selbstwissen zu erwerben (weil sie etwa an einer dauerhaften Fehlfunktion des von

Armstrong postulierten zerebralen Scanners leidet). Shoemaker zufolge können wir uns eine solche ›selbstblinde‹ Person jedoch nicht vorstellen. Jede normal intelligente, begrifflich kompetente und rationale Person, so Shoemaker, wird nämlich wissen, dass sie Äußerungen der Form ›*p* ist zwar der Fall, aber ich glaube nicht, dass *p*‹ tunlichst vermeiden sollte. Denn wer ›Es ist der Fall, dass *p*‹ sagt, bringt damit zugleich zum Ausdruck, dass er glaubt, dass *p*. Wer Letzteres im selben Atemzug verneint, macht sich zwar keines logischen Widerspruchs schuldig – dennoch würde er damit etwas sehr Merkwürdiges sagen. Eine intelligente, begrifflich kompetente und rationale Person wird daher disponiert sein, die Frage ›Glaubst du dass, *p*‹ genau dann zu bejahen, wenn sie auch disponiert ist, die Frage ›Ist *p* der Fall?‹ zu bejahen. Selbstwissen, so scheint es, ist demnach nicht das Ergebnis eines der sinnlichen Wahrnehmung analogen kausalen Mechanismus, sondern stellt sich – solange wir gewisse Rationalitätsstandards nicht unterschreiten – automatisch, ganz ohne das Zutun eines Wahrnehmungsvorgangs ein. Man kann Shoemakers Überlegungen insofern als Plädoyer für die folgende Abwandlung der Phosphoreszenzthese auffassen:

[Phosphoreszenz*] Notwendigerweise gilt: Wenn eine normal intelligente, begrifflich kompetente und rationale Person *S* zum Zeitpunkt *t* glaubt, dass *p*, dann glaubt *S* zu *t* auch, dass sie glaubt, dass *p*.

Der Standardeinwand gegen Shoemakers Überlegung besteht darin, dass sie lediglich zeige, dass sich eine normal intelligente, begrifflich kompetente und rationale Person, die glaubt, dass *p*, *so verhalten wird, als ob* sie glaube, dass sie die Überzeugung, dass *p*, habe. Doch sich so verhalten, als ob man glaube, dass man die Überzeugung, dass *p*, habe, sei etwas anderes als tatsächlich zu glauben, dass man die Überzeugung, dass *p*, habe. Es mag daher sein, dass Shoemaker recht darin hat, dass die Befolgung von Rationalitätsnormen automatisch dazu führt, dass wir ein bestimmtes sprachliches Verhalten an den Tag legen – aber das bedeutet nicht zugleich, dass es automatisch zu Selbstüberzeugungen führt (vgl. Byrne 2018, 47).

6. Variationen der Phosphoreszenz II

Sehen wir uns eine weitere Abwandlung der ursprünglichen Phosphoreszenzthese an, die in der zeitgenössischen Literatur unter dem Titel ›luminosity‹ gehandelt wird:

[Phosphoreszenz**] Notwendigerweise gilt: Wenn eine normal intelligente, begrifflich kompetente und rationale Person *S* zum Zeitpunkt *t* das phänomenale Erlebnis *e* hat, dann ist *S* zu *t* auch in der Lage zu wissen, dass sie *e* hat.

Im Unterschied zu [Phosphoreszenz*], die sich auf Überzeugungen fokussiert, ist die Reichweite von [Phosphoreszenz**] auf phänomenale Erlebnisse beschränkt – wobei hier unter einem phänomenalen Erlebnis entweder eine körperliche Empfindung oder ein sinnlicher Eindruck verstanden werden soll. Die Pointe der Bedingung der normalen Intelligenz, begrifflichen Kompetenz und Rationalität besteht in diesem Fall darin, Fälle wie den des Initiationsrituals auszuschließen. Denn das Subjekt in jener unglücklichen Situation ist aufgrund seiner Befürchtung, mit einem glühenden Schürhaken gebrandmarkt zu werden, zumindest in seiner Rationalität eingeschränkt. Zudem trägt die Bedingung der normalen Intelligenz, begrifflichen Kompetenz und Rationalität der Tatsache Rechnung, dass Tiere oder Kleinkinder Schmerzen empfinden können, jedoch nicht in der Lage sind zu wissen, dass sie Schmerzen haben. Denn Tiere oder Kleinkinder verfügen nicht über die Begriffe, die zur Bildung der entsprechenden Überzeugung notwendig wären. Auch solche Subjekte werden daher von [Phosphoreszenz**] ausgenommen.

Der Umstand, dass im Nachsatz von [Phosphoreszenz**] nicht schlicht von ›wissen‹ die Rede ist, sondern die Wendung ›in der Lage sein zu wissen‹ gewählt wurde, ist ein weiteres Zugeständnis an die Gegnerinnen des cartesianischen Bildes: Die Behauptung besteht nicht darin, dass das bloße Haben eines phänomenalen Erlebnisses bereits hinreichend für entsprechendes Selbstwissen ist – die Behauptung besteht vielmehr darin, dass das Haben eines phänomenalen Erlebnisses zu entsprechendem Selbstwissen führt, sobald das Subjekt alles ihm Mögliche getan hat, um zu entsprechendem Selbstwissen zu gelangen. Dazu zählt etwa, dass es seine Aufmerksamkeit sorgfältig auf sein Erlebnis richtet und im Zuge der Überzeugungsbildung ausschließlich Begriffe verwendet, die ihm nach gewissenhafter Prüfung angemessen erscheinen.

[Phosphoreszenz**] verkörpert den Kern zeitgenössischer Bekanntschaftstheorien, wie sie etwa von Brie Gertler (2012) verteidigt werden. Die Grundidee dieser Theorien besteht darin, dass wir zu unseren eigenen phänomenalen Erlebnissen eine besonders innige epistemische Beziehung unterhalten, die im Großen und Ganzen derjenigen Beziehung gleicht, die Bertrand Russell (1912) als *acquaintance* bezeichnet hatte. Während Russell allerdings unter *acquaintance* eine optimale epistemische Relation verstand, in der – im Sinne des cartesianischen Bildes – der Unterschied zwischen Erscheinung und Wirklichkeit aufgehoben ist, gesteht Gertler die Möglichkeit von Ignoranz und Irrtum durchaus zu (Gertler 2012, 102). Dennoch sei es auf Basis von *acquaintance* möglich, Überzeugungen über die eigenen phänomenalen Erlebnisse zu bilden, die anderen empirischen Überzeugungen hinsichtlich ihres Grades an epistemischer Rechtfertigung weit überlegen seien (ebd., 115).

Timothy Williamson (1996, 557–559) hat ein vieldiskutiertes Argument gegen [Phosphoreszenz**] vorgelegt, das auch als Schlag gegen die Bekanntschaftstheorie aufgefasst werden kann. Im Mittelpunkt des Arguments steht eine Person, die morgens frierend aufwacht, sich jedoch im Laufe des Vormittags kontinuierlich erwärmt, so dass ihr zur Mittagszeit angenehm warm ist. Nehmen wir weiter an, dass diese Person alle in [Phosphoreszenz**] geforderten Bedingungen erfüllt und während des gesamten Zeitraums sorgfältig darauf achtet, wie kalt bzw. warm ihr ist. Sie hat daher zu jedem Zeitpunkt eine Überzeugung darüber, ob ihr kalt ist oder nicht. Man kann Williamsons Argument dann als *reductio ad absurdum* verstehen, die auf dem ›Prinzip der Fehlertoleranz‹ beruht:

> Wenn die beschriebene Person weiß, dass ihr zu einem bestimmten Zeitpunkt kalt ist, dann ist ihr auch eine Millisekunde später kalt (vgl. ebd., 559).

Das Prinzip der Fehlertoleranz leitet sich wiederum aus der von Williamson für plausibel gehaltenen *Safety*-Konzeption von Wissen ab, der zufolge eine wahre Überzeugung nur dann Wissen ist, wenn sie auch in allen hinreichend ähnlichen Situationen wahr gewesen wäre. Da nun der Prozess der Erwärmung im beschriebenen Fall nicht schlagartig erfolgt, sondern sich kontinuierlich über mehrere Stunden hinzieht, stellen zwei durch nur eine Millisekunde getrennte Zeitpunkte während des Geschehens hinreichend ähnliche Situationen im Sinne der *Safety*-Konzeption dar.

Aus [Phosphoreszenz**] und der Beschreibung des Falls folgt nun, dass die Person weiß, dass ihr zum Zeitpunkt t_0, d.h. unmittelbar nach dem Aufwachen, kalt ist. Aus diesem Umstand, zusammen mit dem Prinzip der Fehlertoleranz, folgt wiederum, dass der Person zu t_1, also eine Millisekunde später, kalt ist. Nun greift abermals [Phosphoreszenz**] und führt zu der Behauptung, dass die Person weiß, dass ihr zum Zeitpunkt t_1 kalt ist. Daraus folgt, zusammen mit dem Prinzip der Fehlertoleranz, dass der Person auch eine weitere Millisekunde später, d.h. zu t_2, kalt ist und so

weiter und so fort – bis wir zu dem Resultat gelangen, dass der Person um 12 Uhr mittags kalt ist. Doch dieses Resultat widerspricht der zur Fallbeschreibung gehörenden Annahme, dass der Person zur Mittagszeit angenehm warm ist. [Phosphoreszenz**], so Williamson, kann daher nicht wahr sein (vgl. ebd., 559).

7. Variationen der Unfehlbarkeit I

Tyler Burge (1988) hat eine interessante Abwandlung der Unfehlbarkeitsthese ins Spiel gebracht:

[Unfehlbarkeit*] Notwendigerweise gilt: Wenn Person *S* zum Zeitpunkt *t* urteilt, dass sie denkt, dass *p*, dann denkt *S* zu *t*, dass *p*.

Im Unterschied zur ursprünglichen Unfehlbarkeitsthese ist innerhalb des Vordersatzes das Verb ›glauben‹ durch das Verb ›urteilen‹ ausgetauscht worden und der Platzhalter ›sich in φ befinden‹ durch ›denken, dass *p*‹ – wobei der Kleinbuchstabe ›*p*‹ hier für eine beliebige Proposition steht. Diese Modifikation führt dazu, dass sich keine Situation finden lässt, durch die [Unfehlbarkeit*] entkräftet werden könnte. Der springende Punkt besteht darin, dass das Urteil mit dem Gehalt ›Ich denke, dass *p*‹ nicht existieren kann, solange ich nicht denke, dass *p*. Würde ich zu *t* nicht denken, dass *p*, könnte ich zu *t* auch nicht urteilen, dass ich denke, dass *p*. Oder, noch einmal anders gesagt: *Indem* ich urteile, dass ich denke, dass *p*, denke ich unweigerlich, dass *p*. Burge bezeichnet solche Urteile als ›basic self-knowledge‹ (ebd., 649) und hält sie für geeignet, die besonderen epistemischen Eigenschaften unseres Selbstwissens – seine Direktheit und Sicherheit – zu erklären: Seine Direktheit erkläre sich aus dem Umstand, dass das Zielobjekt des Selbstwissens – der mentale Zustand φ – durch den Vollzug des Urteils mit dem Gehalt ›Ich befinde mich in φ‹ ins Leben gerufen werde; und seine Sicherheit erkläre sich daraus, dass die betreffenden Urteile selbstverifizierend seien, d.h., sie müssen wahr sein, damit sie überhaupt existieren können (vgl. ebd., 649).

Es stellt sich allerdings die Frage, wie erklärungskräftig Burges Hinweis auf Urteile mit dem Gehalt ›Ich denke, dass *p*‹ wirklich ist. Paul Boghossian (1989, 21) hat darauf hingewiesen, dass die meisten Urteile, die gemeinhin als Kandidaten für Selbstwissen gehandelt werden, weder selbstverifizierend sind noch derart, dass sie den mentalen Zustand, auf den sie abzielen, ins Leben rufen. Wenn nun jedoch die meisten unserer Urteile mit dem Gehalt ›Ich befinde mich in φ‹ die von Burge zur Erklärung der Direktheit und Sicherheit unseres Selbstwissens herangezogenen Merkmale gar nicht besitzen, scheint Burges Diagnose kaum etwas wert zu sein.

8. Variationen der Unfehlbarkeit II

David Chalmers (2003) verteidigt eine Version der Unfehlbarkeitsthese, die auf sogenannte ›direkt-phänomenale Urteile‹ zugeschnitten ist. Darunter versteht Chalmers Urteile, in denen sich Personen das Haben eines phänomenalen Erlebnisses mit Hilfe eines sogenannten ›direkt-phänomenalen Begriffs‹ zuschreiben (s. Kap. 27). Bei einem direkt-phänomenalen Begriff handelt es sich wiederum um einen Begriff, der das Erlebnis, auf das er zutrifft, als Komponente enthält. Der direkt-phänomenale Begriff eines Roterlebnisses etwa trifft nicht nur auf dieses Erlebnis zu – das Roterlebnis selbst ist darüber hinaus ein echter Bestandteil dieses Begriffs.

Katalin Balog (2013) zufolge können wir uns einen direkt-phänomenalen Begriff als Resultat ›mentalen Zitierens‹ vorstellen: So ähnlich wie der auf das Wort ›Baum‹ zutreffende Begriff › ›Baum‹ ‹ das Ergebnis einer Operation ist, bei der das Wort ›Baum‹ von Anführungszeichen umschlossen wird, handelt es sich bei dem auf ein bestimmtes Roterlebnis zutreffenden direkt-phänomenalen Begriff um das Ergebnis einer Operation, bei der das Roterlebnis von mentalen Anführungszeichen umschlossen wird. Es bietet sich daher an, direkt-phänomenale Begriffe mit Hilfe der Zeichenfolge ›*◊*‹ zu notieren – wobei die Sternchen für mentale Anführungszeichen und die Raute für das entsprechende Erlebnis stehen. Die von Chalmers verteidigte Unfehlbarkeitsthese lässt sich dann folgendermaßen formulieren:

[Unfehlbarkeit**] Notwendigerweise gilt: Wenn Person *S* zum Zeitpunkt *t* urteilt, dass sie *◊* erlebt, dann ist *S*' Urteil zu *t* wahr.

Wie im Fall der von Burge verteidigten Unfehlbarkeitsthese lässt sich auch für [Unfehlbarkeit**] keine Situation finden, durch die sie entkräftet werden könnte. Und auch der Grund dafür ist ähnlich gelagert wie im vorherigen Fall: Niemand kann ein direkt-phänomenales Urteil, d.h. ein Urteil mit dem Gehalt ›Ich erlebe *◊*‹, fällen, solange das durch die Raute symbolisierte Erlebnis nicht existiert. Denn der im direkt-phänomenalen Urteil verwendete direkt-phänomenale Begriff ›*◊*‹ enthält das Erlebnis, über das geurteilt wird, als Bestandteil. Würde ich dieses Erlebnis nicht haben, könnte ich den Begriff ›*◊*‹ nicht bilden. Und ohne diesen Begriff könnte ich auch das Urteil mit dem Gehalt ›Ich erlebe *◊*‹ nicht fällen. Direkt-phänomenale Urteile sind also in demselben Sinne selbstverifizierend wie Urteile mit dem Gehalt ›Ich denke, dass *p*‹: Ihre Wahrheit ist eine notwendige Bedingung ihrer Existenz.

Wenn man einmal von der Frage absieht, ob es so etwas wie direkt-phänomenale Begriffe überhaupt gibt, weisen Chalmers' Überlegungen ein ähnliches Manko auf wie die Überlegungen Burges: Sie lassen sich nicht auf Urteile über die eigenen Überzeugungen, Wünsche oder Absichten übertragen. Denn bei Überzeugungen, Wünschen oder Absichten handelt es sich typischerweise um nicht-phänomenale Zustände. Es ist uns daher gar nicht möglich, entsprechende direkt-phänomenale Begriffe zu bilden.

9. Die Idee der Transparenz

Im Zuge der Betrachtung der ursprünglichen und modifizierten Versionen der Phosphoreszenz- und Unfehlbarkeitsthesen haben sich mindestens drei Modelle des Selbstwissens herauskristallisiert, die in der zeitgenössischen Diskussion in der einen oder anderen Form vertreten werden: das naturalistische Scannermodell (Armstrong), das rationalistische Modell (Shoemaker) und das Bekanntschaftsmodell (Gertler). Es gibt jedoch mindestens noch ein weiteres Modell des Selbstwissens, das in der zeitgenössischen Diskussion eine wichtige Rolle spielt: das Transparenzmodell (Barz 2012; Byrne 2018). Ein Vorteil des Transparenzmodells wird von seinen Vertretern darin gesehen, dass es – im Gegensatz zum rationalistischen Modell und Bekanntschaftsmodell – eine *alle* Sorten mentaler Zustände abdeckende Erklärung für die Direktheit und Sicherheit von Selbstüberzeugungen liefert, ohne – wie es das naturalistische Scannermodell tut – zu leugnen, dass sich das epistemische Verhältnis, in dem wir zu unserem eigenen Geist stehen, prinzipiell von dem epistemischen Verhältnis unterscheidet, in dem wir zu materiellen Gegenständen in der Außenwelt stehen.

Das Transparenzmodell basiert auf der Beobachtung, dass der Versuch herauszufinden, ob man sich im mentalen Zustand φ befindet, zwangsläufig darauf hinausläuft, dass man den Außenweltsachverhalt in den Blick nimmt, auf den sich φ bezieht. Beispiele für dieses Phänomen lassen sich sowohl im Bereich propositionaler Einstellungen als auch im Bereich phänomenaler Erlebnisse finden. Wenn ich mich etwa frage, ob ich glaube, dass es regnet, suche ich meinen Geist nicht nach der passenden Überzeugung ab – ich frage mich vielmehr, ob es regnet. Beantworte ich diese letzte Frage mit ›ja‹, so weiß ich, dass ich die Überzeugung, dass es regnet, habe (vgl. Evans 1982, 225). Ähnliches gilt auch für den Fall sinnlicher Eindrücke. Wenn ich mich frage, wie mein gegenwärtiger visueller Eindruck beschaffen ist, so wende ich meinen Blick nicht nach innen (um möglicherweise die Rückseite eines aus Sinnesdaten gewobenen Schleiers zu betrachten). Ich nehme vielmehr die vor mir in der Außenwelt befindlichen materiellen Gegenstände ins Visier (vgl. Dretske 1995, 39–63). Man kann sich den Erwerb von Selbstwissen dem Transparenzmodell zufolge daher als ein zweistufiges Verfahren vorstellen: Auf der ersten Stufe fällt das Subjekt ein Urteil über einen Außenweltsachverhalt (›Es regnet‹, ›Dort befindet sich der-und-der Gegenstand‹), um sogleich zu einem Urteil über seine eigenen mentalen Zustände überzugehen (›Ich glaube, dass es regnet‹, ›Ich habe den visuellen Eindruck, dass sich dort der-und-der Gegenstand befindet‹).

Das Transparenzmodell wirft nun jedoch mindestens zwei schwerwiegende Fragen auf. Zum einen setzt es voraus, dass es für jeden mentalen Zustand φ einen Außenweltsachverhalt *p* gibt, auf den er sich bezieht. Im Hinblick auf propositionale Einstellungen und Sinneseindrücke mag das plausibel sein. Aber wie sieht es beispielsweise mit körperlichen Empfindungen aus? Die zweite Frage, die Vertretern des Transparenzmodells beantworten müssen, lautet, wodurch ich eigentlich gerechtfertigt bin, von meinem auf einen Außenweltsachverhalt gerichteten Urteil erster Stufe zu einem höherstufigen Urteil überzugehen, in dem ich mir den Besitz eines bestimmten mentalen Zustands attestiere. Betrachten wir etwa die beiden Propositionen ›Es regnet‹ und ›Ich glaube, dass es regnet‹. Zwischen diesen Propositionen besteht kein inferentieller Zusammenhang: Aus der Tatsache, dass es regnet, folgt nicht, dass eine bestimmte Person glaubt, dass es regnet. Was also verleiht einem Subjekt das epistemische Recht, von der ersten zur zweiten Proposition überzugehen? Diese Frage wird in der Literatur unter dem Titel ›Rätsel der Transparenz‹ verhandelt (vgl. Byrne 2018, 74–96; Barz 2019).

10. Die Autorität der ersten Person

Historisch betrachtet nimmt die Diskussion über die Frage, inwiefern sprachliche Äußerungen der Form ›Ich befinde mich in φ‹ (im Folgenden: ›Selbstzuschreibungen‹) bei ihren Adressatinnen einen Sonderstatus genießen, ihren Ausgang bei Überlegungen Ludwig Wittgensteins (1953). Wittgenstein hatte bestritten, dass es sich bei einer Äußerung wie ›Ich habe Schmerzen‹ um einen Bericht über das Auftreten eines inneren Erlebnisses handelt (vgl. ebd., § 244). Manchmal wird Wittgenstein so missverstanden, als wolle er damit die Existenz innerer Erlebnisse bestreiten. Doch das ist nicht der Punkt, auf den er hinausmöchte. Sein Punkt besteht vielmehr darin, dass es sich bei der Äußerung ›Ich habe Schmerzen‹ um eine besondere Art von Ausdrucksverhalten handle, das vergleichbar mit einem Schrei sei. Genauso wie mit einem Schrei nichts behauptet werde, werde auch mit der Äußerung ›Ich habe Schmerzen‹ nichts behauptet – es werde lediglich der eigene Schmerz *zum Ausdruck gebracht*. Es sei daher unangemessen, solche Äußerungen anzuzweifeln (›Wirklich?‹) oder Gründe einzufordern (›Woher weißt du das?‹).

Der Umstand, dass es unangemessen ist, die Äußerung ›Ich habe Schmerzen‹ in Frage zu stellen, bedeutet Wittgenstein zufolge natürlich nicht, dass wir gezwungen sind, von jeder Person, die den Satz ›Ich habe Schmerzen‹ äußert, anzunehmen, sie habe tatsächlich Schmerzen. Aber etwaige Zweifel beziehen sich dann nicht auf die Wahrheit des Behaupteten – es wird ja gar nichts behauptet! –, sondern auf die Aufrichtigkeit des Gegenübers: Vielleicht will uns unser Gegenüber nur glauben machen, es habe Schmerzen? Allerdings: Solange wir keinen Grund zu der Annahme haben, dass unser Gegenüber unaufrichtig ist, gibt es für uns auch keinen Grund anzunehmen, es habe keine Schmerzen.

Aus diesen, zunächst nur Äußerungen wie ›Ich habe Schmerzen‹ betreffenden Überlegungen Wittgensteins lässt sich die folgende allgemeine These herleiten, die in dieser (oder leicht abgewandelter) Form häufig in der von Wittgenstein beeinflussten Literatur zu finden ist (vgl. Malcolm 1953, 308–309):

[Unbezweifelbarkeit] Notwendigerweise gilt: Wenn eine Person *S* einer anderen Person *H* mitteilt, dass sie (d.h. *S*) sich gegenwärtig im mentalen Zustand φ befindet, dann kann *H* nicht rationalerweise bezweifeln, dass sich *S* in φ befindet, solange *H* annimmt, dass *S* aufrichtig ist und sich nicht versprochen hat oder verwirrt ist.

Aufgrund der Beispiele, die weiter oben im Text bereits im Zusammenhang mit der ursprünglichen Version der Unfehlbarkeitsthese genannt wurden, findet [Unbezweifelbarkeit] in der gegenwärtigen Debatte jedoch nur noch wenige Befürworterinnen. Die meisten zeitgenössischen Philosophinnen bevorzugen eine weniger radikale These, der zufolge in Bezug auf Selbstzuschreibungen lediglich eine Präsumtion der Wahrheit besteht (vgl. Davidson 1984):

[Präsumtion] Notwendigerweise gilt: Wenn eine Person *S* einer anderen Person *H* mitteilt, dass sie (d.h. *S*) sich gegenwärtig im mentalen Zustand φ befindet, und *H* keinerlei Gründe hat, das, was *S* ihm sagt, anzuzweifeln, dann ist *H* darin gerechtfertigt anzunehmen, dass sich *S* in φ befindet – selbst wenn *H* (abgesehen von *S*' Äußerung) über keinerlei Anhaltspunkte verfügt, die die Proposition, dass sich *S* in φ befindet, stützen.

Die zeitgenössische Debatte dreht sich typischerweise um die Frage, wie sich diese Präsumtion erklären lässt. Eine naheliegende Diagnose lautet, dass Hörer voraussetzen, dass Selbstzuschreibungen Überzeugungen zum Ausdruck bringen, die sie mit Hilfe eines privilegierten Zugangs zu ihren eigenen mentalen Zuständen erworben haben. Angesichts der Schwierigkeiten, denen man – wie weiter oben deutlich geworden ist – bei der Erklärung der Direktheit und Sicherheit von Selbstüberzeugungen begegnet, scheint diese Diagnose jedoch unbefriedigend zu sein: Sie erklärt ein rätselhaftes Phänomen im Rekurs auf etwas noch Rätselhafteres. Viele zeitgenössische Philosophen neigen daher zu Diagnosen, die die Autorität der ersten Person ohne Rekurs auf die Idee des privilegierten Zugangs verständlich machen.

Donald Davidson (1984) etwa formuliert ein transzendentales Argument für die Autorität der ersten Person. Er hebt mit der Beobachtung an, dass wir erfolgreich mit anderen Personen kommunizieren: Sie können uns und wir sie verstehen. Dieses wechselseitige Verständnis wäre jedoch nicht möglich, solange die Interpretin nicht annehmen würde, dass der Sprecher weiß, was er mit seinen Äußerungen meint. Da jemand, der weiß, was er mit seinen Äußerungen meint, weiß, was er denkt, läuft die Annahme, dass der Sprecher weiß, was er mit seinen Äußerungen meint, auf die Annahme

hinaus, dass der Sprecher weiß, was er denkt. Davidson zufolge ist die Autorität der ersten Person also eine Bedingung der Möglichkeit sprachlicher Kommunikation im Allgemeinen.

Crispin Wright (1998) hat eine alternative Sicht der Dinge ins Spiel gebracht, in deren Licht die Suche nach einer Erklärung der Autorität der ersten Person ihren Witz verliert. Der von Wright erwogenen (aber letztlich nicht geteilten) Auffassung zufolge handelt es sich bei der Autorität der ersten Person nämlich nicht um die Konsequenz eines tiefer liegenden Phänomens (das durch eine geeignete Erklärung zum Vorschein gebracht werden könnte), sondern schlicht um eine nicht weiter erklärbare primitive Eigenschaft der „Grammatik" psychologischen Vokabulars. Dass Selbstzuschreibungen einen Vertrauensvorschuss genießen, ist dieser Auffassung zufolge eine Auflage, die von jedem Sprecher beachtet werden muss, wenn er denn psychologisches Vokabular kompetent verwenden möchte. Auf die Frage, *warum* oder *mit welchem Recht* diese Auflage besteht, gibt es nicht nur keine Antwort – es ist bereits ein Fehler, diese Frage überhaupt zu stellen.

Literatur

Alston, William: Varieties of Privileged Access. In: American Philosophical Quarterly 8 (1971), 223–241.

Armstrong, David M.: A Materialist Theory of the Mind. London 1968.

Balog, Katalin: Acquaintance and the Mind-Body Problem. In: Simone Gozzano, Christopher S. Hill (Hg.): New Perspectives on Type Identity. The Mental and the Physical. Cambridge 2013, 16–42.

Barz, Wolfgang: Die Transparenz des Geistes. Berlin 2012.

Barz, Wolfgang: Is There Anything to the Authority Thesis? In: Journal of Philosophical Research 43 (2018), 125–143.

Barz, Wolfgang: The Puzzle of Transparency and How to Solve It. In: Canadian Journal of Philosophy 49 (2019), 916–935.

Boghossian, Paul: Content and Self-Knowledge. In: Philosophical Topics 17 (1989), 5–26.

Burge, Tyler: Individualism and Self-Knowledge. In: The Journal of Philosophy 85 (1988), 649–663.

Byrne, Alex: Transparency and Self-Knowledge. Oxford 2018.

Chalmers, David: The Content and Epistemology of Phenomenal Belief. In: Quentin Smith, Aleksander Jokic (Hg.): Consciousness: New Philosophical Perspectives. Oxford 2003, 220–272.

Davidson, Donald: First Person Authority. In: Dialectica 38 (1984), 101–111.

Doppelt, Gerald: Incorrigibility and the Mental. In: Australasian Journal of Philosophy 56 (1978), 3–20.

Dretske, Fred: Naturalizing the Mind. Cambridge MA 1995.

Evans, Gareth: The Varieties of Reference. Oxford 1982.

Gertler, Brie: Renewed Acquaintance. In: Declan Smithies, Daniel Stoljar (Hg.): Introspection and Consciousness. Oxford 2012, 93–127.

Malcolm, Norman: Direct Perception. In: The Philosophical Quarterly 3 (1953), 301–316.

Russell, Bertrand: The Problems of Philosophy. New York 1912.

Ryle, Gilbert: The Concept of Mind. London 1949.

Shoemaker, Sydney: Self-Knowledge and Inner Sense II. The Broad Perceptual Model. In: Philosophy and Phenomenological Research 54 (1994), 271–290.

Williamson, Timothy: Cognitive Homelessness. In: The Journal of Philosophy 93 (1996), 554–573.

Wittgenstein, Ludwig: Philosophische Untersuchungen [1953]. In: Werkausgabe. Bd. 1 Frankfurt a. M. ^{7}1990, 225–580.

Wright, Crispin: Self-Knowledge. The Wittgensteinian Legacy. In: Crispin Wright, Barry C. Smith, Cynthia McDonald (Hg.): Knowing Our Own Minds, Oxford 1998, 13–45.