

HS: Der Geist der Tiere  
Sommersemester 2004  
Dozenten: Prof. Dominik Perler & Dr. Markus Wild

**Intentionalität –  
Zwei Ansätze zur Begriffsklärung**

Oktober 2004

Jan Koernicke,  
8. Fachsemester

Samariterstr. 1  
10247 Berlin  
[jan.koernicke@gmx.de](mailto:jan.koernicke@gmx.de)

## Inhalt

<b>1. Einleitung</b>	<b>3</b>
<b>2. Intentionalität als weiter Begriff</b>	
<b>2.1. Intentionalität als Zuschreibung</b>	<b>4</b>
<b>2.2. Kritik</b>	<b>7</b>
<b>3. Intentionalität als Merkmal von kognitiven Agenten</b>	
<b>3.1. Reaktive Agenten</b>	<b>10</b>
<b>3.2. Kognitive Agenten</b>	<b>12</b>
<b>3.3. Naturalisierte Intentionalität</b>	<b>15</b>
<b>3.4. Kritik</b>	<b>16</b>
<b>4. Fazit</b>	<b>19</b>
<b>5. Bibliographie</b>	<b>20</b>

## 1. Einleitung

In der theoretischen Philosophie sind nur wenige Begriffe so unbestimmt wie das Konzept von Intentionalität. Verstanden wird Intentionalität zumeist als Eigenschaft mentaler Zustände, sich auf etwas zu beziehen – also als wesentliches Merkmal von Überzeugungen, Hoffnungen, Wünschen etc. Tatsächlich ist es plausibel zu sagen, dass derartige mentale Zustände immer einen Bezug haben müssen. Intentionalität wird daher auch oft übersetzt als *aboutness*, *Gerichtet-sein* oder als eine Art von *Gedankeninhalt*. Allerdings sind dies nur Umschreibungen; eine anerkannte Definition oder ein praktisches Kriterium gibt es nicht.

Dies liegt offensichtlich an der besonderen Art dieser Eigenschaft. Franz Brentano, der den Begriff im 20. Jahrhundert zuerst wieder aufgreift, meint, dass jedes psychische Phänomen durch eine *intentionale Inexistenz* ausgezeichnet wäre<sup>1</sup>, womit er intentionale Phänomene von physikalischen abgrenzen wollte. Bis heute werden in dem Zusammenhang die verschiedensten Fragen diskutiert: ‚Gibt es intentionale Objekte?‘, ‚Kann Intentionalität naturalisiert – also auf eine materialistische Basis zurückgeführt – werden?‘, ‚Ist Intentionalität ein Merkmal aller mentalen Zustände?‘ etc<sup>2</sup>.

Im folgenden sollen zwei verschiedene Ansätze zum Verständnis des Begriffes vorgestellt und untersucht werden. Zum einen eine sehr weite Definition von Intentionalität, die sich an Daniel C. Dennett orientiert und zum anderen ein Versuch, der über eine Klassifizierung von Agenten funktioniert, indem er diese unter bestimmten Umständen als intentional definiert – ähnlich Fred Dretskes Ansatz zur Verhaltensklärung. Insbesondere sollen dabei nicht nur Menschen, sondern auch Tiere und Maschinen als potentielle Träger von Intentionalität betrachtet werden.

---

<sup>1</sup> In [Brentano, 1874]

<sup>2</sup> Siehe dazu bspw. [Jacob, 2003]

## 2. Intentionalität als weiter Begriff

### 2.1. Intentionalität als Zuschreibung

Von Daniel C. Dennett stammt das Konzept einer *intentional stance*<sup>3</sup>, welche Intentionalität als eine von außen zugeschriebene Eigenschaft betrachtet. Dennett schlägt vor, man solle Systeme, deren Verhalten einen Rückschluss auf Intentionalität erlaubt, als intentionale Systeme behandeln. Dieser Rückschluss erfolgt, indem man prüft, ob die Zuschreibung von Intentionalität hilfreich ist, um das Verhalten eines Systems zu erklären und – wichtiger noch – sein zukünftiges Verhalten vorauszusagen:

„The intentional strategy consists of treating the objects whose behaviour you want to predict as a rational agent with beliefs and desires and other mental stages exhibiting what Brentano and others call *intentionality*. (...) Then I will argue that any object – or as I shall say, any *system* – whose behaviour is well predicted by this strategy is in the fullest sense of word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behaviour is reliably and voluminously predictable via the intentional strategy.“<sup>4</sup>

Für Dennetts Strategie spielen die eigentlichen Eigenschaften des Systems, dem Intentionalität zugeschrieben wird, keine Rolle – entscheidend ist nicht, ob ein System *intentional ist*, sondern ob man ihm Intentionalität *zuschreibt*; er räumt ein, dass man auch Thermostaten und Pflanzen als intentionale Systeme behandeln kann. Das einzige Kriterium ist für ihn, ob sich damit das Verhalten des Systems im allgemeinen korrekt voraussagen lässt.

Alternativ zur *intentional stance*, aber in einem sehr ähnlichen Sinn, kann man Intentionalität auch als Verhaltensdisposition verstehen. Indem man nämlich einem System bestimmte Intentionen unterstellt, schreibt man ihm damit gleichzeitig zu, sich in bestimmten Situationen auf eine bestimmte Art zu verhalten. Wenn ich z.B. jemandem den Wunsch nach Erdbeereis (und keine dem entgegen stehenden Überzeugungen) zuschreibe, so muss ich annehmen, dass er die Disposition hat, sich Erdbeereis zu kaufen, wenn sich für ihn eine

---

<sup>3</sup> Siehe dazu sein gleichnamiges Buch [Dennett, 1987]

<sup>4</sup> [Dennett, 1987], S. 15

Möglichkeit ergibt. Der Unterschied zwischen Dennetts *intentional stance* – also einer Art Einstellung – und Verhaltensdispositionen ist also im wesentlichen ein technischer; wobei der Begriff der Verhaltensdisposition den Vorteil besitzt, dass er das wesentliche Element von Dennetts Strategie – nämlich die Verhaltensvoraussage – bereits als Begriff impliziert.

Die konkrete Zuschreibung von Intentionalität in diesem Sinne geschieht über die Zuschreibung von Wünschen und Zielen, die ein rationales System verfolgen würde. Dennett formuliert eine kurze Anleitung, wie seine Strategie im Detail funktionieren soll:

„First you decide to treat the object whose behaviour is to predict as a rational agent; then you figure out what belief that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this intentional agent will act to further goals in the light of its beliefs.“<sup>5</sup>

Was zunächst sehr gewagt erscheint, – denn wieso sollte man Pflanzen oder gar Magneten mentale Zustände zuschreiben, die auch noch die Eigenschaft haben, sich auf andere Dinge zu beziehen? – erweist sich doch als eine sehr nützliche Strategie: Tatsächlich scheinen wir sie unbewusst immer wieder anzuwenden; auch im täglichen Sprachgebrauch schlägt sich diese Zuschreibung von Wünschen und Zielen gegenüber Maschinen oder noch einfacheren Gegenständen nieder. (,Der Blitz will ins Wasser', ,Der Videorecorder glaubt, es wäre acht Uhr', ,Die Sonnenblume will zum Licht') Der Vorteil dieser Art von Erklärung liegt ganz einfach darin, dass es oft die einzige Erklärung ist, die wir haben. Während man das Verhalten eines Thermostates anhand von einfachen physikalischen Regeln auch auf andere Weise erklären kann (Dennett verwendet hierzu die *physical* und die *design stance*), so fällt dies bei vielen anderen Systemen schwer. Um aber das Verhalten dieser komplexeren Systeme auch ohne besondere Fachkenntnisse erklären und vorhersagen zu können, ist die Zuschreibung von Intentionalität oft die einzige praktikable Möglichkeit.

Zweifellos liefert die Strategie der *intentional stance* oder das Konzept von Intentionalität als Verhaltensdisposition ein Verständnis von Intentionalität, welches nicht der philosophisch (z.Zt.) gängigen Bedeutung des Begriffs entspricht. Doch ist dies – gerade in Anbetracht der großen Probleme, mit der das Konzept Intentionalität verbunden ist – nicht zwingend ein Argument gegen den Ansatz. Unbestreitbar erhält man auf diese Weise einen eindeutigen und v.a. funktionierenden Begriff (Dennett spricht von einem „extraordinarily powerful tool

---

<sup>5</sup> [Dennett, 1987], S. 17

in prediction“<sup>6</sup>), der alles erfasst, was gemeinhin als Intentionalität verstanden wird. Und dass er darüber hinaus noch weiter greift und auch Systeme als intentional auszeichnet, welchen man diese Eigenschaft prinzipiell eher absprechen würde, ist womöglich nicht diesem speziellen Ansatz verschuldet, sondern vielleicht ein grundsätzliches Problem: Die Tatsache, dass es bisher kein überzeugend funktionierendes Konzept von Intentionalität gibt, welches nur die Systeme erfasst, die man intuitiv als intentional bezeichnen würde, kann auch als Indiz dafür gedeutet werden, dass dieser Anspruch prinzipiell gar nicht erfüllbar ist und das Verständnis von Intentionalität nicht funktionieren kann, wenn es nicht auf Systeme ausgeweitet wird, die bislang als nicht-intentional galten.

Ein weiteres Argument, das ähnlich funktioniert und die Verwendung eines derartig weiten Begriffs von Intentionalität plausibel macht, erläutert Dennett explizit:

„It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather when we discover some objects for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system“<sup>7</sup>

Es erscheint durchaus plausibel, dass das Funktionieren der *intentional stance* als Erklärungsstrategie ein Hinweis auf das tatsächliche Vorliegen von mentalen Repräsentationen sein kann, wenn es sonst keine plausible Erklärung gibt. Warum sollte man nur aus dem Vorhandensein von mentalen Repräsentationen auf Intentionalität schließen, nicht jedoch aus dem Auftreten von nur intentional erklärbarem Verhalten auf das Vorliegen von mentalen Repräsentationen? Dennetts Argument ist nicht von der Hand zu weisen – zumal er damit deutlich macht, dass sein Konzept keine ontologische Neutralität bzgl. Intentionalität fordert, sondern von ihm durchaus im Einklang mit einer materialistischen Theorie von Intentionalität gesehen wird.

Die konkrete Frage nach potentiellen Kandidaten für Intentionalität hat Dennett damit klar beantwortet: Grundsätzlich kann jedes System als intentional angesehen werden – die Zuschreibung erfolgt nach dem Maßstab des Beobachters. Als Grenze funktioniert dabei der Anspruch, die angelegte Theorie so komplex wie nötig und so einfach wie möglich zu

---

<sup>6</sup> [Dennett, 1987], S. 24

<sup>7</sup> [Dennett, 1987], S. 32

halten. Sollte sich das Verhalten eines Systems also einfacher beschreiben lassen als durch die Zuschreibung von Intentionen (z.B. durch die Erklärung physikalischer Zusammenhänge; Dennetts *physical stance*), so macht es wenig Sinn, Intentionalität zuzuschreiben.

Praktisch gelten nach einem derartigem Ansatz außer Menschen auch eine große Klasse von Tieren als intentionale Wesen. Im Bereich der Automaten & Maschinen wird diese Zuschreibung wohl stark variieren – während die jeweiligen Programmierer und Designer das Verhalten des Systems völlig ohne Bezug auf Intentionen erklären können, so wird die Mehrzahl der Beobachter z.B. bei einem Fußballspiel der Robocup<sup>8</sup>-Roboter kaum auf die Zuschreibung von Wünschen, Zielen und Überzeugungen verzichten können, wenn sie das Verhalten der putzigen Maschinen erklären will.

## 2.2. Kritik

Eine Theorie, die Intentionalität unter einem derartig weiten Begriff verstehen will, sieht sich natürlich verschiedenen kritischen Argumenten ausgesetzt. Zwei schwache Einwände lassen sich einfach ausräumen bzw. klären, zwei weitere, tiefergehende Einwände gegen ein solches Verständnis wiegen jedoch schwerer.

Zuerst kann eine Zuschreibung von Intentionalität im weiten Sinn der *intentional stance*, bzw. der Verhaltensdisposition immer nur eine sehr konkrete Zuschreibung sein: Zugeschrieben werden kann immer nur eine ganz spezielle Überzeugung, ein konkreter Wunsch oder ein bestimmtes Ziel. Es ist nicht möglich, einem System aufgrund seines Verhaltens direkt eine allgemeine Intentionalität zuzuschreiben. Statt dessen muss immer der ‚Umweg‘ über eine konkrete innere Einstellung genommen werden, welche dann das allgemeine Vorliegen von Intentionalität impliziert. Genau genommen geht es also nicht um das Zuschreiben von Intentionalität, sondern zuerst um das Zuschreiben von Wünschen, Überzeugungen, Hoffnungen etc. Der Bezug auf Intentionalität als Eigenschaft des Systems kann nur indirekt erfolgen.

Dies mag auf den ersten Blick als Manko erscheinen, lässt sich aber mit dem Verweis auf die Zuschreibung von anderen Eigenschaften leicht ausräumen: Ob ein Objekt beispielsweise farbig ist, lässt sich auch nicht direkt ‚ermitteln‘ – auch hier muss der (implizite) Schritt über die Zuschreibung einer konkreten Farbe gemacht werden, um ein Objekt als farbig zu qualifizieren. Offensichtlich ist diese Art von Eigenschafts-Zuschreibung also ein Verfahren, das durchaus gängig und unkontrovers ist. Es wäre also wenig plausibel, in diesem Fall mehr zu fordern.

---

<sup>8</sup> siehe <http://www.robocup.org/> [25. Oktober 2004]

Ein zweites, ebenfalls nicht besonders starkes Argument, betrifft die Menge von intentionalen Zuständen, die den Systemen zugeschrieben werden können. Dennett geht explizit davon aus, dass lediglich *beliefs* und *desires* zugeschrieben werden sollen (s. Zitat S. 5). Doch ist es durchaus fraglich, ob alle Arten innerer Zustände, die sich auf etwas beziehen, als Überzeugungen oder Wünsche charakterisiert werden können; eine Zuordnung von Befürchtungen beispielsweise scheint nur schwer möglich.

Doch ist die Frage, ob man Befürchtungen als Überzeugungen oder Wünsche behandeln kann, kaum Inhalt einer Theorie über Intentionalität. Zweifellos würde Dennett problemlos erklären können, wie er mit solchen Fällen verfahren will – naheliegend wäre beispielsweise, den belief-Begriff so zu erweitern, dass er nicht nur ‚direkte‘ Überzeugungen erfasst.

Stärker wiegen zwei andere Argumente gegen die ‚weite Intentionalität‘. Zum einen ist eine derartige Intentionalität genaugenommen eine ‚Als-ob-Intentionalität‘. Dennetts Vorschlag besteht darin, einem System Intentionalität zuzuschreiben, wenn es sich so verhält *als ob* es die Eigenschaft tatsächlich besitzt. Auf den zweiten Blick erkennt man, dass dieser Vorschlag als Definition nicht funktioniert, da er zirkulär funktioniert. Eine Erklärung der Art

„*Intentional*“ wenn „*Verhalten, als ob Intentional*“

ist nicht in der Lage, zu erklären, was unter ‚intentional‘ verstanden werden soll. Hinter einem derartig weiten Intentionalitäts-Verständnis versteckt sich ein implizit vorausgesetztes Verständnis des fraglichen Begriffs.

Fraglich ist allerdings, ob Dennett mit seinem Konzept der *intentional stance* tatsächlich eine Definition von Intentionalität liefern will. Als Strategie zur Erklärung bestimmter Phänomene und als Anleitung zur Klassifizierung von Systemen in einem größeren Kontext kann die *intentional stance*, die von Dennett zusammen mit der *physical* und *design stance* in erster Linie zur Charakterisierung von *True Believers* benutzt wird, durchaus nützlich und ergiebig sein. Dass sie darüber hinaus als Definition für Intentionalität den Makel der Zirkularität aufweist, ist ein Problem – aber womöglich nicht Dennetts Problem.

Ein anderer, ebenfalls gewichtiger Kritikpunkt ist die Betrachterabhängigkeit der Zuschreibungen. Eine Eigenschaft, über deren Zuschreibung verschiedene Individuen verschiedener Meinung sein können, mag grundsätzlich nicht problematisch sein – Begriffe wie ‚schön‘, ‚angenehm‘ oder ‚geschmacklos‘ beziehen ihren ganzen Reiz aus ihrer Relativität – aber in wissenschaftlichen Kontexten sind sie in der Regel wenig hilfreich.

Dennett ist sich der Problematik bewusst:

„Would it not be intolerable to hold that some artifact or creature or person was a believer [*und damit intentionales System – J.K.*] from the point of view from one observer, but not a believer at all from the point of view of another, cleverer observer?“<sup>9</sup>

Mit einer kurzen Entgegnung tut er diese Bedenken jedoch ab:

„The decision to adopt the intentional stance is free, but the facts about the success or failure of the stance, were one to adopt it, are perfectly objective.“<sup>10</sup>

Offensichtlich ist Dennett der Ansicht, dass man Objektivität dadurch erlangt, dass die Fakten und das Wissen, auf deren Grundlage zwei verschiedene Beobachter zwei verschiedene Zuschreibungen vornehmen, miteinander abgeglichen werden. Zwei Beobachter, die mit denselben Informationen ausgestattet, ein System klassifizieren, können gemäß dieser Annahme nicht mehr zu zwei unterschiedlichen Schlüssen kommen – womit das Kriterium Objektivität erlangt hätte.

Doch Zweifel sind angebracht: Würden sich beide Beobachter mit demselben Faktenwissen (Oder benötigt es mehr als reines Faktenwissen; und was könnte das sein?) immer auf dieselbe Zuschreibung verständigen? Und ist dieses ‚Abgleichen‘ von Wissensständen überhaupt ein zulässiges Verfahren, wenn die Herangehensweise zuvor explizit als betrachterabhängige Zuschreibung eingeführt wurde? Es drängt sich der Verdacht auf, dass die Objektivität hier durch die Hintertür hereingeholt werden soll, nachdem man sie am Haupteingang abgewiesen hat.

Offensichtlich bietet der Ansatz, Intentionalität als einen weiten Begriff zu verstehen, prinzipiell ein ordentliches Handwerkzeug, das diese mysteriöse Eigenschaft aus einem gewissen Unschärfbereich herausholt und ein anwendbares Kriterium liefert. Doch bleiben im Detail Probleme, die zwar prinzipiell nicht völlig unüberwindbar scheinen – das Konzept aber liefert zumindest in dieser Form nur schwerlich eine wissenschaftlich zufriedenstellende Definition.

---

<sup>9</sup> [Dennett, 1987], Seite 23f.

<sup>10</sup> [Dennett, 1987], Seite 24

### 3. Intentionalität als Merkmal von kognitiven Agenten

#### 3.1. Reaktive Agenten

Eine engerer Begriff von Intentionalität sieht Agenten als potentielle Träger von Intentionalität. Als Agenten seien ganz basal verschiedenste physische oder virtuelle Entitäten verstanden, die weitestgehend autonom funktionieren und mit ihrer Umwelt interagieren. Diese Anforderungen erfüllen Menschen, Tiere, Pflanzen, eine Vielzahl von Maschinen und auch bestimmte Software-Programme<sup>11</sup>. Der Begriff des Agenten erfasst damit alle Systeme, denen man im weitesten Sinne Handlungsfähigkeit zugestehen kann. Es scheint wenig erfolgversprechend, Entitäten, die nicht autonom sind oder welche nicht mit ihrer Umwelt interagieren, als Kandidaten für Intentionalität zu betrachten. Niemand würde bspw. auf die Idee kommen, einem Arm oder einem Stein Intentionalität und damit ein gewisses geistiges Innenleben zuzuschreiben.

Was unterscheidet nun also intentionale Agenten von nicht-intentionalen Agenten? Ruth Milikan liefert in ihrem Aufsatz *Varieties of Purposive Behaviour* folgenden Ansatz:

„Das Wort ‚intentional‘ wird von Philosophen verwendet, um auf Dinge Bezug zu nehmen, die von anderen Dingen handeln z.B. der Satz ‚Paris ist schön‘ und ein Stadtplan von Paris. (...) Äußere Dinge, die Intentionalität manifestieren, etwa Sätze, geographische Darstellungen, Schaubilder, Karten, Straßenzeichen, Musiknoten und darstellende Gemälde werden „Repräsentationen“ genannt. Eine vorherrschende Theorie, der ich zustimme, schlägt vor, dass innere intentionale Dinge, etwa Überzeugungen, Hoffnungen, Wünsche und Intentionen, in ähnlicher Weise Repräsentationen sind. Allgemeiner ausgedrückt: Alle Kognitionen sind innere Repräsentationen – d.h. innere Modelle im abstraktesten mathematischen Sinn – dessen, wovon sie handeln. Der Unterschied zwischen rein biologischen Zwecken und intentionalen Zwecken liegt darin, dass im zweiten Fall die biologischen Zwecke des Lebewesens durch die Herstellung und Verwendung von inneren Repräsentationen implementiert

---

<sup>11</sup> Der Begriff des (*Software-*) *Agenten* ist in der Informatik weit verbreitet und ein breites, aktuelles Forschungsfeld, siehe z.B. [Brenner, 1998] oder [Ferber, 2001].

werden – Repräsentationen von der Umwelt und/ oder Repräsentationen von den Zielen des Lebewesens.“<sup>12</sup>

Die Diskussion um die Theorie der inneren oder ‚mentalen‘ Repräsentationen soll an dieser Stelle nicht geführt werden. Eine sehr plausible Einführung und Verteidigung des Konzeptes findet sich in Tim Cranes *The mechanical mind*<sup>13</sup>.

Doch inwiefern ist der Verweis auf mentale Repräsentationen hilfreich bei der Suche nach einem Kriterium für Intentionalität? Es scheint wenig sinnvoll, ‚intentional‘ als Eigenschaft von Repräsentationen zu definieren und anschließend festzustellen, dass mentale Repräsentationen intentional funktionieren. Das Problem scheint lediglich verschoben.

Doch ganz so vergeblich ist der Ansatz nicht. Fragt man nämlich danach, wie ein Agent funktionieren würde, der nicht intentional ist, also nicht über ein System mentaler Repräsentationen und damit nicht über Überzeugungen, Wünsche, Hoffnungen oder ähnliches verfügt, so erhält man eine durchaus vielversprechende Antwort: Nach Abzug aller Formen von mentalen/ inneren Repräsentationen bleiben rein reaktive Systeme übrig: Da diese Systeme keinerlei Informationen speichern; sich keinerlei ‚Vorstellung‘ von den Dingen in der Welt machen können, können sie nur auf direkte Wahrnehmungen reagieren. Sie zeichnen sich damit durch eine besonders einfache Funktionsweise aus - auf einen bestimmten Input reagieren sie zuverlässig immer wieder mit einem bestimmten Output. Auf diese Agenten wirkt eine ‚imperative Kraft des Gegebenen‘ – ihr Verhalten ist nicht abhängig von Stimmungen, Zielen, Meinungen oder Wünschen; sie funktionieren lediglich, indem sie auf bestimmte Bedingungen - auch *trigger* genannt – warten und reagieren, sobald diese Bedingungen erfüllt sind. Pflanzen, die ihre Blüten nach dem Sonnenstand öffnen und schließen oder Thermostate, die über Bimetallstreifen auf Temperaturveränderungen reagieren, sind klassische Vertreter solcher *reaktiver Agenten*.

Doch ist es nicht das alleinige Auftreten von solch rein reaktiven Verhaltensweisen, die einen Agenten als nicht-intentionalen, reaktiven Agenten qualifiziert. Auch bei Menschen – zweifellos intentionale Agenten – lässt sich derartig reaktives, unbedachtes, reflexartiges Verhalten feststellen; bspw. das Schließen der Augen bei schneller Annäherung eines Gegenstandes. Das Kriterium für einen echten reaktiven Agenten ist, dass er *nur* über derartige reaktive Verhaltensweisen verfügt. Diese Agenten zeichnen sich im allgemeinen auch dadurch aus, dass ihre Wahrnehmung auf wenige ‚Kanäle‘ beschränkt ist. So scheinen Ameisen, bei denen der Geruchssinn eine herausragende, alle anderen Sinne dominierende Rolle spielt, oder Amphibien, bei denen optische Wahrnehmungen immer wieder und unter

---

<sup>12</sup> [Perler & Wild, 2005], Seite (?)

<sup>13</sup> [Crane, 2003]

allen Umständen ein bestimmtes Verhalten hervorruft, als Systeme zu funktionieren, die ohne eine mentale/ innere Repräsentation ihrer Außenwelt funktionieren.

Die Frage, ob ein Agent über mentale Repräsentationen und damit über Intentionalität verfügt, oder ob dies eben nicht der Fall ist, kann also durchaus auf empirisch beobachtbares Verhalten zurückgeführt werden. Neben dem ‚imperativen Zwang des Gegebenen‘, unter dem Systeme stehen, die nicht-intentional sind, ergibt sich eine Reihe weiterer Eigenschaften, die diese Systeme aufweisen – bspw. sind sie nicht lernfähig, da sie dazu das Erlebte verarbeiten und repräsentieren müssten. Die Frage, ob ein Tier oder eine Maschine über die Merkmale eines rein reaktiven Agenten verfügt, oder ob es als intentionaler, *kognitiver Agent* angesehen werden muss, kann also – zumindest bis zu einer bestimmten Gewissheit – empirisch untersucht und geklärt werden.

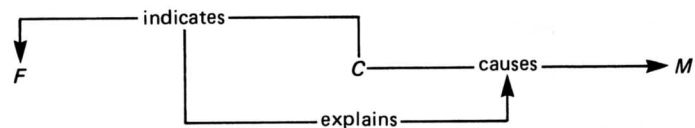
### **3.2. Kognitive Agenten**

Unabhängig von der empirischen Frage nach den Merkmalen eines reaktiven Agenten gibt es prinzipielle Fragen, die bezüglich potentiell kognitiver Agenten gestellt werden müssen. Was genau heißt es, mentale Repräsentationen zu haben? Wie funktionieren derartige Agenten? Und sollen wirklich alle Agenten, die über innere Repräsentationen verfügen, als intentional betrachtet werden?

Ein einfaches Modell für die Funktionsweise von kognitiven Agenten ist die sogenannte BDI-Architektur. Diese erklärt das Verhalten von Systemen anhand einer Menge von *beliefs* und *desires*, über welche die Agenten in Form von mentalen Repräsentationen verfügen. Beliefs, also Überzeugungen oder Meinungen, sind dabei Repräsentationen, von denen ein Agent ‚glaubt‘, dass sie aktuell vorliegen – es sind gewissermaßen seine Annahmen über den momentanen Zustand der Welt. Außerdem können beliefs auch Annahmen über die Funktionsweisen der Welt repräsentieren; z.B. könnten die Sätze ‚Der Stein liegt auf dem Boden‘ und ‚Wenn ich den Stein hochhebe und los lasse, fällt er herunter‘ beliefs sein. Als desires, also Wünsche, werden diejenigen Repräsentationen verstanden, welche einen Zustand repräsentieren, den der Agent – aus welchen Gründen auch immer – für erstrebenswert hält. Die Verhaltensweisen eines BDI-Agenten lassen sich nun mit Bezug auf seine beliefs und desires erklären: Da er bestimmte Überzeugungen über die Welt besitzt und

gewisse Ziele hat, unternimmt er Aktionen, von denen er meint, dass sie zur Erfüllung seiner Ziele führen.<sup>14</sup>

Ein ähnliches Konzept von kognitiven Agenten entwickelt Fred Dretske in *Explaining Behavior*<sup>15</sup>. Er entwirft ein Modell, welches außer beliefs und desires auch *causes*, *facts*, *movement* und *reinforcement* enthält. Eine einfache Form seines Konzepts stellt er wie folgt dar:



“M [movement – J.K.] is being produced by an internal C [cause]. Furthermore ... this causal relationship between C and M, if its going to be explained by something like the meaning of C, will have to be explained by the fact that C indicates, or has the function of indicating, how things stand elsewhere in the world. It will not be enough merely to have a C that indicates F [fact] about M – the fact about C that explains, or helps explain, *why* it causes M. What needs to be done, then, is to show how the existence of one relationship, the relationship underlying C’s semantic character, can explain the existence of another relationship, the causal relationship (between C and M) comprising the behavior in question.”<sup>16</sup>

Dretske geht es nicht nur darum, darzustellen, dass C – ein belief – ein bestimmtes Verhalten (M-ovement) verursacht. Er betont, dass dieser Zusammenhang nur dadurch erklärt werden kann, dass C einen bestimmten Fakt in der Welt repräsentiert<sup>17</sup>. Das Modell, das er insgesamt entwickelt, ist sehr detailliert – so skizziert er u.a., wie ein solches System lernen kann.

Interessant für die Betrachtung von kognitiven Agenten ist bei Dretske ein besonderer Punkt:

---

<sup>14</sup> Einen formalen Ansatz der BDI-Architektur und einen Vorschlag zur prinzipiellen Implementation in Software-Agenten formulieren Anand Rao und Michael Georgeff in ihrem Aufsatz *BDI Agents: From Theory to Practice* [Rao & Georgeff, 1995].

<sup>15</sup> [Dretske, 1988]

<sup>16</sup> [Dretske, 1988], S. 83f.

<sup>17</sup> Wobei Dretske streng genommen eine scharfe Unterscheidung zwischen *indicate* und *represent* macht; diese ist hier aber nicht wesentlich.

“Though instruments and machines don’t have beliefs and desires, much less do things *because* of what they believe and desire, they nevertheless *do* things.”<sup>18</sup>

Wie kommt es, dass obiges Modell für Maschinen nicht gelten soll? Und warum sollen Maschinen prinzipiell keine beliefs und desires haben? Für Dretske ist dies ein grundlegender Punkt: Während innere Repräsentationen bei Menschen generell immer einen Fakt der Welt repräsentieren, ist dies bei Maschinen nicht der Fall. Die Repräsentationen der Maschinen erhalten ihre *Bedeutung* von uns, von ihren Konstrukteuren; daher sind diese Repräsentationen – nach Dretske – eben keine beliefs oder desires. Bezugnehmend auf die Funktionsweise der obigen Darstellung in Maschinen erklärt er:

“If anyone or anything is responsible for C’s causing M ... it is we, its creators. So *we* caused C to cause M. We did so, however, because of some fact C.”<sup>19</sup>

Dretske unterscheidet, ebenso wie John Searle, der in diesem Zusammenhang von ‚Intrinsischer Intentionalität‘ spricht<sup>20</sup>, zwischen inneren Repräsentationen im Menschen, die sich zweifelsohne auf äußere Zustände beziehen, und Repräsentationen von Maschinen, die sich eben nicht ‚einfach so‘ auf Fakten in der Welt beziehen, sondern die diese Bedeutung erst durch uns bekommen.

Eine derartige Unterscheidung ist wenig plausibel. Es mag unter gewissen Umständen Sinn machen, zu unterscheiden, *woher* der Zusammenhang zwischen innerer Repräsentation und äußerem Zustand kommt – wobei diese Frage selbst bei Menschen keineswegs einfach zu beantworten ist. Aber an der Art dieser Repräsentationen, an der simplen Tatsache, *dass* sie sich auf einen Fakt beziehen – und damit an dem wesentlichen Merkmal der Intentionalität – kann kaum gezweifelt werden. Searle wie Dretske scheinen unbedingt ein Unterscheidungskriterium für eine Eigenschaft zu suchen, das mit der eigentlichen Eigenschaft herzlich wenig zu tun hat. Für die Frage *ob* jemand gebräunte Haut hat, ist es völlig irrelevant, ob er diese vom spanischen Strand mit nach Hause gebracht oder im Sonnenstudio um die Ecke erlegen hat. Die Unterscheidung von mentalen Repräsentationen, die sich nur deshalb auf etwas beziehen, weil das so vom Konstrukteur des Systems geplant

---

<sup>18</sup> [Dretske, 1988], S. 85

<sup>19</sup> [Dretske, 1988], S. 86

<sup>20</sup> [Searle, 1980]

war, und Repräsentationen, die sich aus anderen – schwer zu benennenden Gründen – auf etwas beziehen, ändert nichts daran, *dass* sie sich auf etwas beziehen.

Unabhängig davon liefert Dretske ein sehr durchdachtes Modell für das Funktionieren von kognitiven Agenten, welches die einfache BDI-Architektur aufnimmt und weiter verfeinert. Doch ist ein zu konkretes Modell insofern problematisch, als dass es sich als zu speziell erweisen kann, um bestimmte Systeme zu erfassen. Daher scheint es ratsam, zur Charakterisierung dessen, was einen kognitiven Agenten im wesentlichen ausmacht, zum allgemeinsten möglichen Mittel zu greifen. Und dies ist grundsätzlich – wie von Milikan (s. Zitat S. 10) angeführt – das Vorliegen von inneren, bzw. mentalen Repräsentationen. Genauer: Das Vorliegen eines Systems von Repräsentationen, welches in gewisser Hinsicht inferentiell funktioniert, bzw. fähig ist, zumindest minimale Schlussfolgerungen aus diesem Repräsentationssystem zu ziehen.

Die Frage nach Systemen, die lediglich über einzelne Repräsentationen – also über kein System – verfügen, bzw. Systemen, die nicht in der Lage sind, aus einem vorhandenem System neue Konsequenzen zu ziehen, ist interessant, aber wahrscheinlich ohne besondere Relevanz, da diese Fähigkeiten einzeln einem System keinerlei Nutzen bringen und als solche – evolutionär bedingt – kaum auftreten und v.a. keinerlei Auswirkungen auf die Funktionsweise des Systems haben.

### **3.3. Naturalisierte Intentionalität**

Grundlegend für das dargestellte Konzept von Intentionalität ist dessen Naturalisierung. Das System von Repräsentationen muss materialistisch manifestiert sein, da es sonst keinerlei Kraft besitzt. Dies liegt zum einen daran, dass es wenig Sinn macht, (kausal wirksame) Repräsentationen ohne materialistische Verpflichtung anzunehmen. Wenn die inneren Einstellungen, Wünsche, Überzeugungen etc. als Erklärung für etwas benutzt werden sollen, bspw. zur Erklärung von Verhalten, müssen sie tatsächlich *vorhanden* sein. Ein bestimmtes Verhalten oder auch die Eigenschaft, Intentionalität zu besitzen, kommen nicht *irgendwie* in die Welt; sie müssen eine materialistische Basis haben, um kausal wirksam zu sein. Ohne diese ontologische Verpflichtung kann das Konzept nicht funktionieren.

Darüber hinaus hätte das aufgezeigte Konzept kaum Erklärungskraft, wenn es nicht auf einer naturalisierten Basis funktionieren würde. Ohne diese Basis wäre die Rede von Repräsentationen beliebig, da unklar ist, wie diese Repräsentationen funktionieren sollen.

Mentale Repräsentationen sind in einem nicht naturalistischen Verständnis ein leeres Konstrukt ohne jede Erklärungskraft.

Im Gegensatz dazu ist das Konzept eines weiten Begriffes von Intentionalität (siehe Abschnitt 2) nicht auf eine ontologische Verpflichtung angewiesen. Da es sich dabei lediglich um eine Zuschreibung von außen handelt, ist es nicht nötig anzunehmen, dass diese Zuschreibungen auch materialistisch grundiert sind.

Am Beispiel eines Schachcomputers lässt sich der Unterschied zwischen dem weiten Begriff und dem engeren Konzept von Intentionalität auch in Hinsicht auf die Notwendigkeit einer ontologischen Aussage verdeutlichen:

Denkbar ist zum einen ein Schachcomputer, der seine Zug-Entscheidungen aus einer großen Datenbank heraus trifft, in der alle möglichen Schachstellungen zusammen mit jeweils einem bestimmten, ‚optimalen‘ Zug gespeichert sind. In jeder möglichen Situation ‚schaut‘ das Programm in dieser Datenbank nach und führt den entsprechenden Zug aus.

Eine anderes Programm bewertet die jeweils aktuelle Situation nach einem bestimmten Schlüssel (nach Figuren und Stellung) mit einem numerischen Wert. Jede Situation, die durch einen Spielzug erreicht werden kann, wird erneut bewertet. Dann wird der Zug, der den höchsten Wert erreicht, ausgeführt. Es geht für dieses Programm also darum, einen möglichst hohen Wert zu erreichen (welcher im Spiel einem Schach-Matt entspricht).

Beide Programme können nach der *intentional stance* als intentional behandelt werden: Nach dem zuerst dargestellten, weiten Intentionalitäts-Konzept gibt es keinen Unterschied zwischen den beiden Computerprogrammen.

In der Unterscheidung nach reaktiven und kognitiven Agenten wird jedoch das erste Programm als reaktiv, das zweite als kognitiv eingeordnet. Die Datenbankabfragen des ersten System liefern eine eindeutige Handlungsanweisung – ohne dass die jeweilige Situation selbst im System repräsentiert wird. Das zweite Computerprogramm hingegen nimmt Bewertungen vor, die sich auf den Spielstand beziehen. Es verarbeitet die äußeren Fakten – und muss sie dazu repräsentieren. Die Differenzen zwischen den verschiedenen möglichen Stellungen, welche jeweils repräsentiert werden, führt zu einer Entscheidung. Dieses kognitive System repräsentiert also äußere Situationen; es repräsentiert sie konkret im Arbeitsspeicher eines Computers in Form von Bits und Bytes; Nullen und Einsen.

### 3.4. Kritik

Ein grundlegendes Problem des Konzeptes ist die Frage, wie man es im Detail verstehen will. Wie dargestellt besteht für ein Modell, das wie Dretskes sehr detailliert erklärt, wie Intentionalität funktioniert, die Gefahr, dass es nicht zutrifft. Umso detaillierter das Modell ist, umso größer ist die Wahrscheinlichkeit, dass es sich empirisch als falsch erweisen wird. Die Alternative, ein grobes Modell, das lediglich eine Art von inferentiellem Repräsentationsmodell fordert, ist zwar allgemein genug, aber deutlich weniger plausibel. Ein zu abstraktes Modell besitzt weniger Erklärungskraft.

Diesem grundsätzlichen Dilemma kann man nicht entgehen. Wenn man ein Konzept entwickeln will, das einerseits Erklärungskraft besitzen, andererseits aber grundsätzlich empirisch nachprüfbar sein soll, besteht immer die Gefahr, dass die Erklärung zu speziell wird und sich daher in einzelnen Punkten als falsch erweist. Formuliert man die Theorie jedoch zu allgemein, geht der eigentliche Anspruch verloren. Hier ist abzuwägen, in welcher Form die Theorie am besten funktionieren kann – gegen das konkrete Konzept ist das Problem jedoch kaum ein geeignetes Argument.

Ein ähnliches Problem stellt die Unterscheidung von reaktiven und kognitiven Agenten dar. Einerseits besteht der Anspruch, dass die Klassifizierung empirisch erfolgen kann – andererseits gibt es ohne Zweifel eine Reihe von Fällen, in denen eine solche Klassifizierung zumindest schwer fällt. Die Merkmale und Unterschiede von reaktiven und kognitiven Agenten scheinen theoretisch klar, praktisch müssen sie das aber keineswegs sein. Was genau ist ein Repräsentationssystem? Wann ist es inferentiell? Und was sind eigentlich mentale Repräsentationen?

Im Einzelfall kann es hier sicher zu Problemen kommen, die eine Entscheidung problematisch machen. Aber auch dies ist ein Problem, das weniger die konkrete Theorie sondern eher die Art der Definition betrifft. Bei nahezu allen praktisch-empirischen Klassifizierungen gibt es Grenzfälle; ob dies nun die astronomische Definition von Planeten oder die biologische Definition von Vögeln betrifft. Wenn man den (hohen) Anspruch einer empirisch anwendbaren Definition hat, lässt sich dieses Problem kaum vermeiden; um zu einer Lösung zu gelangen, müsste man die Theorie auf ein deutlich weniger gewinnbringendes Niveau reduzieren.

Ein letzter möglicher Einwand, der an dieser Stelle diskutiert werden soll, richtet sich grundsätzlich gegen das Kriterium reaktiv oder kognitiv als Unterscheidungskriterium von intentionalen und nicht-intentionalen Agenten. Es scheint, als könne ein bedeutend einfacheres Kriterium dasselbe leisten: Eine Unterscheidung bzgl. der *Reaktionsbandbreite* der Agenten. Reaktive Agenten – als nicht-intentionale Systeme – verfügen offenbar in jeder

Situation immer nur über eine mögliche Reaktion. Kognitive Agenten – als intentionale Systeme – hingegen können (auf Grund ihrer inneren, intentionalen Zustände) auf eine Situation mit verschiedenen Aktionen reagieren; entsprechend ihren inneren intentionalen Repräsentationen. Ein Kriterium, das lediglich anhand der Reaktionsbandbreite eine Unterscheidung trifft, scheint dementsprechend einfacher: Wenn der Agent auf eine Situation nur mit einer einzigen Aktion reagieren kann, ist er nicht-intentional, wenn er aber über eine ganze Bandbreite von verschiedenen Reaktionen verfügt, dann muss er auch über Intentionalität verfügen.

Ist es aber tatsächlich so, dass der kognitive Agent auf eine konkrete Situation mit verschiedenen Aktionen reagieren kann? Muss nicht eine materialistische Position<sup>21</sup> einräumen, dass die Handlung des kognitiven Agenten durch seine konkreten Zustände – inklusive seiner Überzeugungen, Wünsche etc. – keineswegs ganz verschieden ausfallen kann, sondern dass ein konkreter Agent in einer bestimmten Situation mit den eigenen augenblicklichen Zuständen nur auf eine bestimmte Weise reagieren kann?

Die Frage ist der Diskussion um das Verständnis von Willensfreiheit zumindest sehr ähnlich, wenn nicht mit dieser identisch. Offensichtlich muss ein Ansatz, der Intentionalität auf diese Weise erklären will, die Diskussion um die Willensfreiheit zumindest teilweise übernehmen. Dadurch verliert das Konzept allerdings die einfache Plausibilität, die es gegenüber dem ursprünglichen Ansatz besaß. Die mögliche Reaktionsbandbreite der Agenten ist als Erklärung für Intentionalität zumindest solange nicht überzeugend, wie die Diskussion um Willensfreiheit nicht zu einem Ende geführt wurde.

---

<sup>21</sup> Für alle anderen hat das Konzept wahrscheinlich schon auf Seite 15 verloren.

#### 4. Fazit

Beide dargestellten Ansätze sind nicht ohne Probleme. Allerdings sind die Einwände nicht so schwerwiegend, dass man eines der Modelle vollständig zurückweisen müsste. Sowohl die Theorie eines weiten Begriffes von Intentionalität als auch die engere, an kognitive und reaktive Agenten angelegte Theorie, sind funktional und ergiebig.

Auch sind beide Ansätze insofern kompatibel, als dass sie mit verschiedenen Begriffen von Intentionalität arbeiten – zum einen als Zuschreibung, zum anderen als materialistisch grundierte ontologische Verpflichtung. Welches Verständnis und welches Kriterium für Intentionalität man anlegen will, bleibt freigestellt; keine der beiden Theorien verfügt über einen nachprüfbaren, empirischen Vorteil. Allerdings scheint das verbreitete Verständnis von Intentionalität in den meisten Fällen schwer verträglich mit dem Konzept, das Dennett vorschlägt: Vor allem die Relativität, die das Konzept durch seine Abhängigkeit vom Betrachter bekommt, erscheint intuitiv nicht überzeugend.

Für die konkrete Frage danach, welche Agenten als intentional bezeichnet werden können und sollten, ist die erste Definition daher auch nur bedingt geeignet; je nach Betrachter könnte quasi allen Agenten – inklusive Pflanzen – oder aber keinen – wenn nämlich selbst das Handeln von Menschen auf deren ‚Design‘ zurückgeführt wird – Intentionalität zugeschrieben werden.

Betrachtet man Intentionalität jedoch als Eigenschaft kognitiver Agenten, welche über ein System von mentalen Repräsentationen verfügen, so wird man neben Menschen auch eine Reihe höherer (Säuge-) Tierarten und bestimmte Maschinen bzw. deren Programme als intentional ansehen. Bemerkenswerter Weise ist dies dann (bisher) lediglich im Fall der Computerprogramme eine objektiv nachprüfbare Tatsache; einzig die Repräsentationen im Speicher der Maschinen sind klar als solche identifizierbar.

Der Klärung des Begriffes der Intentionalität ist mehr als reiner Selbstzweck. Für die Probleme des Selbstbewusstseins, der Sprachphilosophie, der künstlichen Intelligenz und das große Feld der Kognitionswissenschaften stellt Intentionalität in der einen oder anderen Form ein wichtiges Thema dar. Daher scheint es durchaus ratsam, zuerst zu einer klaren und anerkannten Definition dieses schwierigen Begriffes zu kommen.

## 5. Bibliographie

- Brenner, Walter : *Intelligente Softwareagenten : Grundlagen und Anwendungen*. Berlin [u.a.] : Springer, 1998.
- Brentano, Franz: *Psychologie vom empirischen Standpunkte*. Leipzig: Duncker & Humblot, 1874
- Crane, Tim : *The mechanical mind : a philosophical introduction to minds, machines, and mental representation*. - 2. ed. . - London [u.a.] : Routledge, 2003.
- Dennett, Daniel Clement : *The intentional stance*. - Cambridge, Mass. : MIT Pr., 1987
- Dretske, Fred I.: *Explaining behaviour : reasons in a world of causes*. - Cambridge, Mass. [u.a.] : MIT Pr., 1988.
- Ferber, Jacques : *Les systèmes multi-agents <dt.> Multiagentensysteme : eine Einführung in die Verteilte Künstliche Intelligenz*. München [u.a.] : Addison-Wesley, 2001.
- Jacob, Pierre, "*Intentionality*", The Stanford Encyclopedia of Philosophy (Fall 2003 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2003/entries/intentionality/>
- Perler, Dominik & Wild, Markus (Hrsg.): *Der Geist der Tiere. Philosophische Texte zu einer aktuellen Diskussion*, F.a.M.: Suhrkamp 2005.
- Rao , Anand S. and Georgeff, Michael P.. *BDI agents: From theory to practice*. In: Proceedings of the First International Conference on Multi-Agent Systems, pages 312-319, 1995.
- Searle, John: *Intrinsic Intentionality*. In: Behavioral and Brain Sciences 3: 450-456, 1980.