Vorgelegt an der Alpen-Adria Universität Klagenfurt

Selbstwahrnehmung auf dem Weg von natürlichen in künstliche Organismen

Proseminar Bewusstseinspsychologie
Ass.-Prof. i.R. Dr. Gottfried Süssenbacher

von
Markus Echterhoff
Matrikelnr.: 0560785
Klagenfurt
2. März 2010

Inhaltsverzeichnis

1 Abstract

Diese Arbeit setzt sich mit Selbstwahrnehmung im Sinne der bewussten Wahrnehmung der eigenen körperlichen Existenz und Kognition in evolutionärem, sowie technischem Kontext auseinander. Ausgehend von der Grundannahme, dass organisatorisch geschlossene, strukturell offene Systeme entsprechend [Maturana and Varela, 1987] die Systemeigenschaften über Strukturveränderungen hinaus erhalten, schließe ich von der Funktionsweise natürlich-evolutionär entwickelter kognitiver Apparate sozialer Wesen auf die bestehende Möglichkeit der erfolgreichen Erzeugung künstlicher Wesen, die über Selbstwahrnehmung verfügen. Zur Unterstützung dieser Folgerung dienen aktuelle Forschungen zur künstlichen Metakognition und künstlichen Sozialevolution.

2 Einleitung

Nehmen wir als Vorraussetzung für das Wahrnehmen des eigenen Selbst westlich-philosophisch, dualistisch ein Subjekt und ein Objekt der Erkenntnis an, so stoßen wir rasch an eine Grenze, denn wenn das Objekt der Erkenntnis das Subjekt selbst ist, gibt es keinen externen Standpunkt mehr von dem aus das Objekt betrachtet werden kann [Smith, 1986]. Das subjektive Empfinden verhält sich dieser Erkenntnis jedoch zum trotz: *Ich* reflektiere über *meine* Wahrnehmung, *meine* Gedanken und *meine* Reflektionen. Es scheint als gäbe es eine zentrale kognitive Einheit, die über ihre eigenen Prozesse reflektiert.

Dieses Gefühl von Einheit ist eine Illusion, wie Experimente mit Split-Brain-Patienten (Callosotomie) zeigen. Vielmehr entstehe das „normale Bewusstsein“ aus dem Zusammenspiel der linken und rechten Hemisphäre, die sich durchaus auch widersprechen können, was zu inneren Konflikten führen könne. Aufgrund der Dominanz der linken Hemisphäre über die rechte seien diese Konflikte jedoch nicht offensichtlich. [Zaidel et al.,] Ich wage zu vermuten, dass die Trennung weiterer Verbindungen, wären sie ebenso leicht zu lokalisieren und operieren, zur Erkenntnis führen würde, dass Bewusstsein in großen Teilen des Gehirns gleichzeitig existiert und unsere Selbstwahrnehmung ein Produkt dieser Bewusstseinsprozesse ist. *Wie* diese Wahrnehmung von Selbst entsteht, ist nicht Teil dieser Arbeit, hier verfolge ich die Fragen *warum* sie entsteht und ob es möglich ist sie in künstlichen Medien zu erzeugen.

Dass Menschen nicht die einzigen Wesen dieses Planeten sind, die sich selbst wahrnehmen können, gilt heute als gesichert. So konnten sich etwa Elephanten, Delphine, Rabenvögel (Corvidae) und Menschenaffen im weiteren Sinn (Hominoidea) selbst im Spiegel identifizieren [Sadedin and Paperin, 2009], Ratten über ihre eigene Wahrnehmung reflektieren [Foote and Crystal, 2007] und Schimpansen wird eine *Theory of Mind* anerkannt [Kim and Lipson, 2009], dennoch wird einer Maschine in der Regel ungern jegwede Form von Bewusstsein zugesprochen und Maschinenbewusstsein blieb bisher Science-Fiction. Moderne Forschungsgebiete, wie Artificial Intelligence, Cognitive Science und Artificial Life, sind bemüht unser Verständnis von Bewusstsein zu erweitern. In dieser Arbeit verfolge ich einen evolutionären Zugang zum Verständnis von Bewusstsein.

Zuerst gehe ich auf die Begriffe *Evolution* und *Sinn* ein, deren tiefes Verständnis Vorraussetzung für die folgenden drei Kapitel ist, in denen ich zunächst eine evolutionäre Theorie zur Entwicklung von Selbstwahrnehmung in der Natur erläutere und danach diese Theorie, über die Unterscheidung zwischen Organisation und Struktur dynamischer Systeme als Brücke, in die Möglichkeit zur künstlichen Evolution menschenähnlicher Selbstwahrnehmung überführe.

3 Evolution und Sinn

Ich unterscheide in dieser Arbeit zwischen *natürlicher Evolution*, als die von Charles Darwin beschriebene Evolution biologischer Lebensformen, *künstlicher Evolution* als einen dazu analogen, von Menschen gezielt eingeleiteten Prozess und schließlich *Evolution* bzw. *evolutionärer Prozess* im Allgemeinen als die Systementwicklung eines dissipativ-selbstorganisierenden Systems fern vom Gleichgewicht (s. [Echterhoff, 2010]). Diese Unterscheidung führe ich einzig zum Zweck eines leichteren sprachlichen Verständnisses durch, es handelt sich aus abstrakter Sicht um gleiche Prozesse. In dieser Arbeit wird Evolution als ein Zufallsprozess verstanden, der zusammen mit einem einzigen Gesetz das heutige Universum inklusive der Menschheit und Technik hervorgebracht hat. Dieses Gesetz nannte Charles Darwin *survival of the fittest*. Dawkins erkannte in dieser Beschreibung einen Spezialfall eines allgemeineren Gesetzes: *survival of the stable*, mit dem sich die natürliche Evolution auch vor dem ersten Leben auf der Erde erklären lässt [Dawkins, 2006]. Es handelt sich auch hier um sprachliche Details, denn wir können *fit* mit *geeignet* oder *tauglich*

übersetzen und das, was am geeignetsten ist, ist auch jenes, welches stabil bleibt. Es scheint, als bestünden diese stabilen Dinge *für* etwas, als erfüllten sie einen Zweck, als hätten sie einen Sinn. Evolution als Zufallsprozess jedoch kennt keinen Sinn. Dennoch erscheint alles, was sie hervorbringt sinnvoll. So sinnvoll, das manche Menschen dahinter einen Schöpfer vermuten.

Um die Existenz bestehender Dinge zu begründen, spezifiziere ich also den Begriff *Sinn*. Etwas macht Sinn, genau dann, wenn es der Förderung von Stabilität (bzw. Tauglichkeit) eines Objekts eines evolutionären Prozesses fördert. Um diese Definition zu verstehen, muss verstanden werden, dass unser Gehirn selbst in einem evolutionärem Prozess entstanden ist und seine Organisation auf das Überleben im Sinne der Evolution hin ausgerichtet ist; unabhängig von der Einheit der Selektion[1]. In Retrospektive hat daher jedes Ereignis, das zur Existenz eines Objekts der Evolution geführt hat, *Sinn* gemacht. Es gilt zu beachten, dass diese Beobachtung ausschließlich in Retrospektive gemacht werden kann: Evolution kennt keint keine Planung. Planung selbst ist etwas, das sich in der natürlichen Evolution als stabil erwiesen hat und darum so exzessiv vom menschlichen Gehirn betrieben wird.

4 Natürliche Selbstwahrnehmung

Das Wort *Selbstwahrnehmung* bezeichnet in dieser Arbeit zum einen die Fähigkeit den eigenen Körper zu erfahren, über durchgeführte Handlungen und Sinneswahrnehmungen zu reflektieren und zum anderen bezeichnet es das Phänomen, sich selbst als den Denker von Gedanken zu erfahren, im Gegensatz dazu, über die eigenen Gedanken vollständig definiert zu werden. Die buddhistische Lehre kennt als sechstes Sinnesorgan das *Geistorgan* mit dem die Gedanken wahrgenommen werden (s. [Hanh, 1999]), so dass Selbstwahrnehmung ebenfalls eine Sinneswahrnehmung ist, gleichwertig in allen Aspekten außer dem Bereich der Wahrnehmung,

[1] Als Selektionseinheit wird eine abstrakte Einheit eines Evolutionsprozesses bezeichnet, die sich gegenüber anderen Einheiten durchsetzt, oder nicht. Seit dem ersten Bekanntwerden der Evolutionstheorie streiten Wissenschaftler um das Ausmaß dieser Einheit. Werden Gruppen (zB. eine Spezies) selektiert? Oder vielmehr einzelne Individuen? Oder sind es in Wirklichkeit die Gene dieser Individuen? (s.a. [Dawkins, 2006]) Ich verstehe Evolution als allgemeingültigen Algorithmus, der durch die Komplexität der bestehenden Einheiten (Gene, Individuuen und Gruppen) einen mehrschichtigen Prozess bildet, in dem all diese Einheiten selektiert werden, manche längerfristiger als andere. Der Begriff *Selektion* wird hier hauptsächlich aus Konformitätsgründen verwendet, der Leser ignoriere für diese Arbeit den aktiven, intentionalen Charakter des Wortes.

denn das Geistorgan ist nach Innen gerichtet und nimmt Gedanken und Gefühle wahr.

Für dieses Kapitel stelle ich die Folgende Frage: *Warum existiert Selbstwahrnehmung?* Oder: *Welchen Sinn hat Selbstwahrnehmung?*

Im Kontext zum vorherigen Kapitel muss Selbstwahrnehmung einen evolutionären Vorteil schaffen, wiederum unabhängig von der Einheit der Selektion. Wenn dem so ist, warum existieren dann so viele Lebewesen, die sich nicht selbst wahrnehmen können? Warum ist der Mensch scheinbar das einzige Lebewesen, das ein *Geistorgan* besitzt? Wir können gefahrlos sagen, dass der Mensch zumindest in dem von ihm wahrgenommenen Universum das erste und bislang einzige solche Wesen ist. Wenn Evolution als Zufallsprozess mit Selektion verstanden wird, dann gibt es für jedes Phänomen ein erstes Auftreten und bis auf Widerspruch können wir daher annehmen, dass der Mensch das erste Auftreten einer Mutation (dieses Ausmaßes) ist, die wir als das *Geistorgan* bezeichnen können. Die, in der Einleitung erwähnten Studien bestätigten zwar, dass Menschen nicht die einzigen Wesen sind, die zu Formen der Selbstwahrnehmung fähig sind, aber dennoch setzt sich menschliche Selbstwahrnehmung von der anderer Wesen offensichtlich ab. Um dies zu erklären, beschreibe ich eine fiktive Evolution.

Bei der Betrachtung eines evolutionären Prozesses muss man sich für einen Punkt entscheiden, bei dem man einsteigt. Im Rahmen dieser Arbeit reicht eine Welt aus, die von Wesen besiedelt ist, die nicht in der Lage sind, sich auf irgendeine Weise wahrzunehmen. Diese Wesen können als biologische Roboter verstanden werden. Ein solcher Biobot bestehe aus Motorik (Muskeln), einer schützende Außengrenze, einem Energieversorgungssystem zur Aufnahme und Weiterleitung von Energie an die Motoren, Sensoreinheiten zur Außenwahrnehmung, regenerative und reproduktive Vorrichtungen, sowie einer rudimentären Steuerungseinheit (Nervensystem) und gegebenen Falls ein Skelett. Das Nervensystem sei fähig, Fortbewegung und Greifaktionen zu koordinieren und auf Sinnesreize mit genetisch festgelegten Handlungen zu reagieren. Weniger imaginative Leser dürfen etwa an ein Insekt denken und dabei einen nicht allzugroßen Fehler machen. Wie Evolution vom Urknall weg die Existenz eines solchen Biobots erklärt, ist bereits diskutiert worden und kann im ohnehin empfehlenswerten Buch „The Selfish Gene“ von Richard Dawkins nachgelesen werden.

Von diesem Startpunkt aus laufe die Evolution weiter. Angenommen der Zufallsprozess führe zum ersten unter allen diesen Wesen, das

Schmerz als negativ empfinden kann, der ihm etwa durch die Penetration seiner Außengrenze zugefügt wird. Dieses Wesen würde seine Artgenossen überleben, die sich nicht davon stören ließen, von einem scharfen Gegenstand langsam durchtrennt zu werden. Bald würden also schmerzempfindliche Wesen dominieren. Der Zufall erzeuge eine weitere Mutation und in der Folge das erste Wesen, das sich Dinge merken kann. Dieses Wesen hat deutlich höhere Überlebenschancen, da es eine einmal überlebte Gefahrensituation im zweiten Anlauf erinnert und dadurch präemptiv tätig werden kann. Entsprechend [Humphrey, 1999] kann sich hier bereits eine Form von Selbstwahrnehmung (original: *consciousness*, dt.: *Bewusstsein*) bilden. Denkbar wäre etwa die Beziehung zwischen visueller Wahrnehmung des eigenen Körpers und Schmerzrezeptoren in der Außenhaut. Wann immer etwas visuell die Außenhaut penetriert, würde es sich eine innere Repräsentation speichern, die es davon abhielte sich in die visuelle Nähe solcher Gegenstände zu begeben, es könnte also Reiz-Reaktions-Muster erlernen (Konditionierung), was ihm einen dramatischen Vorteil verschaffen würde.

Nun besitzen also nach einiger Zeit viele dieser Wesen ein Gedächtnis, weil die, die keines haben nach und nach aussterben. Diese Wesen können, wie erwähnt, eine aus Sinneswahrnehmung und Gedächtnis hervorgehende Vorstellung von sich selbst besitzen. Der Zufall will es nun, dass sich eines dieser Wesen an etwas erinnert, das gar nicht geschehen ist. Es entwickelt die Fähigkeit der Imagination[2] und damit auch die der Planung. Solche Mutationen, die eine Selbstrepräsentation in der Imagination zulassen, setzen sich durch, da sie es erlauben, in Form von Simulationen, begrenzt die Zukunft vorherzusagen und somit noch früher präemptiv tätig zu werden und weitere Gefahren für das Überleben abzuwenden.

Nach einer genügend langer Zeitspanne wird sich die Mutation weitesgehend durchgesetzt haben und die Welt ist bevölkert von planenden Wesen. Es wurde bereits mehrfach gezeigt, dass Kooperation lernender Wesen ein emergentes Phänomen ist und unter vielfältigen Bedingungen auch künstlich erzeugt werden kann [Andras, 2009], [Phelps et al.,

[2] Ein populistisches Beispiel eines solchen evolutionären Vorteils wird im Film „The Invention of Lying“ (2009) gezeigt, in dem in einer fiktiven, modernen, gegenwärtigen Welt die gesamte Menschheit nichts als die Wahrheit denkt und spricht und auch nichts verschweigt, bis ein einziger Mensch die Fähigkeit entwickelt, an etwas anderes als die Wahrheit zu denken und auch zu sprechen. Lässt man sich von den Logikfehlern nicht davon abhalten das Gedankenexperiment wirken zu lassen, bekommt man einen guten Eindruck davon, was geschieht, wenn eine Mutation einer Selektionseinheit einen großen evolutionären Vorteil verschafft.

2009] und [Iizuka et al., 2009]. Wir können also davon ausgehen, dass unsere Wesen, im Überlebenskampf gegen ihre Umgebung, beginnen kooperierende Verbunde zu bilden. Während unsere Wesen zuvor lediglich mit der Außenwelt um Energieressourcen rangen, emergiert nun ebenfalls ein Selektionsprozess innerhalb der kooperierenden Gruppen, sowie zwischen eben solchen. Es werden bei Nahrungsknappheit diejenigen Gruppenmitglieder überleben, die am sich am klügsten[3] verhalten. Körperliche Charakteristika sind nicht länger ausreichend. So können Allianzen innerhalb der Gruppe gebildet werden um bestimmte Individuen (die ebenfalls Nahrungsmittel verbrauchen) auszugrenzen, selbst wenn diese körperlich überlegen sind. Der Social Brain Hypothesis [Dunbar, 1998] nach entsteht menschenähnliche Kognition durch eine solche Sozialevolution. Unsere Wesen würden nun also evolvieren, Verhalten und Absichten ihrer Mitwesen zu erlernen, solange Selektionsdruck vorherrscht. Fügen wir ohne Begrenzung der Allgemeinheit die Existenz von Geschlecht und Sex hinzu, so wird zusätzlich zur Nahrung auch um Partner gebuhlt werden, was zu einer Beschleunigung des evolutionären Prozesses führt[4], da der Selektionsdruck sich erhöht, denn es kann sich nicht jeder fortpflanzen, selbst dann, wenn er überlebt. Durch die schrittweise Verlagerung der Selektion vom Konflikt mit der Umwelt (wie etwa durch die Erfindung von Werkzeugen), auf den Konflikt untereinander, werden entsprechend [Dunbar, 1998] komplexere Gehirne selektiert. Geben wir unseren Biobots also genügend Zeit und Selektionsdruck, dann besteht die Möglichkeit, dass sie Verhaltensweisen aufweisen, die wir zumindest von sozialen Tieren kennen, wenn nicht sogar von Menschen.

Im gesamten Prozess der natürlichen Evolution scheint die Serialisierung der Gehirnfunktionen in einen einzigen Bewusstseinsstrom, der das Ich-Erleben hervorbringt, nicht nötig zu sein. Wir kommt es dann dazu, dass sich Menschen als die Denker ihrer Gedanken erfahren, wenn es keinen evolutionären Vorteil verschafft? Ich stelle die Hypothese auf, dass dieses Phänomen eine Folge der Kombination von Planungsfähigkeit und Sozialentwicklung ist, statt eine eigenständige Entwicklung. Planungsfähigkeit erfordert ein Gedankenobjekt, das den eigenen Körper repräsentiert und in einem kausalen Universum existiert. Kausalität erfordert eine „geht voraus“-Relation, die wiederum Zeitempfinden benötigt, das durch

[3] Klugheit ist ebenso zu verstehen, wie *Sinn* zuvor spezifiziert wurde.

[4] Und dadurch auch erklärt, warum wir uns durch Sex fortpflanzen: Die Mutation, die Sex hervorbrachte, sorgte für eine Entwicklungsbeschleunigung, der andere Fortpflanzungsmethoden nachstanden, wodurch sich Spezies (bzw. Gene) durchgesetzt haben, die sich mehrgeschlechtlich fortpflanzen.

eine entsprechende Serialisierung paralleler Kognition erreicht wird. Hier liegt also eine primitive Form von linear-kausalem Ich-Erleben vor. Sozialevolution hingegen fördert Empathie und das Vorhersagen von anderen, die ähnlich dem imaginierten Geistobjekt des eigenen Körpers sind und doch anders. Dieses *anders* konstituiert sich entsprechend den Selektionsfaktoren als das, was wir allgemein als *Persönlichkeit* bezeichnen. Durch Interaktion mit diesen Persönlichkeiten bildet sich rückwirkend das Eigenbild, mit dem das Geistobjekt des eigenen Körpers bestückt wird um noch genauere Vorhersagen über zukünftige Interaktionen treffen zu können. Es bildet sich eine *Theory of Mind*.

Dass wir als Menschen uns kognitiv so unterschiedlich von den Tieren wiederfinden, dass es scheint als seien wir nicht Teil der Natur, führe ich darauf zurück, dass menschliche Selbstwahrnehmung durch ein selbstverstärkendes, katalytisches Netzwerk von Phänomenen beschleunigt wird, das die Lücke zwischen Tieren und Menschen größer werden lässt. Ein solches katalytisches Phänomen, das bereits genannt wurde ist Sex. Sex scheint nicht notwendig für die Entwicklung von Gehirnen, aber beschleunigt diese so maßgeblich, dass der Großteil der Wesen der Erde inzwischen Sex zur Fortpflanzung verwendet. Dazu zählen weiter komplexe Sprache, Werkzeuggebrauch und Mathematik, die sich gegenseitig beinflussen und enorm zum evolutionären Vorteil der menschlichen Spezies beigetragen haben. Aufgrund ihrer besonderen Bedeutung seien die Erfindungen des Computers und des Internets genannt, die unsere Evolution weiter explodieren lassen und die Entwicklung unserer Gehirne beeinflussen.

5 Systemorganisation und Systemstruktur

Dass es Formen tierischen Bewusstseins gibt, die denen des Menschen nicht vollständig unähnlich sind, scheint für viele Menschen leichter zu akzeptieren, als dass es eine Maschine geben könnte, die Bewusstsein besitzt. Dieses Kapitel soll den Unterschied ausräumen. Dafür verwende ich die Begriffe *Organisation* und *Struktur* im Bezug auf Systeme. Ein *System* kann in diesem Rahmen als Menge beliebiger Einheiten (*Systemkomponenten*), die miteinander in Verbindung stehen und zusammen ein Ganzes der Erkenntnis bilden, verstanden werden.

> „Unter Organisation sind die Relationen zu verstehen, die zwischen den Bestandteilen von etwas gegeben sein müssen, da-

> mit es als Mitglied einer bestimmten Klasse erkannt wird.“ [Maturana and Varela, 1987]
>
> „Unter der Struktur von etwas werden die Bestandteile und die Relationen verstanden, die in konkreter Weise eine bestimmte Einheit konstituieren und ihre Organisation verwirklichen.“ [Maturana and Varela, 1987]

Diese Definitionen sind leicht nachvollziehbar: Denken wir an ein beliebiges Spiel, seiner Bekanntheit wegen etwa „Mensch, Ärgere Dich Nicht“, dann besteht dieses Spiel aus den Spielfiguren, von denen sich eine fixe Anzahl von allen anderen unterscheidet, um sie einem Spieler eindeutig zuordnen zu können, einem Spielbrett, einem Zufallszahlengenerator (idR. ein Würfel) und einem sehr einfachen und festen Regelwerk. Das ist die *Organisation* des Spiels. Ob die Spielfiguren groß oder klein sind, aus Metall oder aus Plastik, ändert nichts daran, dass wir es als „Mensch, Ärgere Dich Nicht“ erkennen. Wir könnten dieses Spiel ebenso mit virtuellen Figuren am Computer spielen, es bliebe das gleiche. Wir verändern die *Struktur*, aber erkennen es an seiner *Organisation* weiterhin als „Mensch, Ärgere Dich Nicht“.

Analog muss eine Replika der *Organisation* eines menschlichen Gehirns menschliche Eigenschaften aufweisen, unabhängig von der konkreten *Struktur*. Ob wir natürliche oder künstliche Neuronen verwenden und an natürliche oder künstliche Sinnesorgane und motorische Systeme anschließen, führt dieser Logik folgend zum gleichen Ergebnis, unter der Vorraussetzung, dass die Organisation auch wirklich gleich bleibt - sämtliche, an der Organisation beteiligten, Charakteristika der Komponenten also identisch sind. Nach [Dennett, 1998] sind sämtliche biologischen Prozesse mechanisch reproduzierbar, Dennet hielt es jedoch für möglich, wenn auch unwahrscheinlich, dass die Geschwindigkeit und Kompaktheit biochemischer Nervensysteme die technische Realisierung von Gehirnen begrenzt. Eine von ihm vorgeschlagene und von Kevin Warwick experimentell verfolgte[5] Möglichkeit ist es, Biochemie und Robotik zu Cyborgs zu kombinieren, so dass die schnellen und kompakten biologischen Nerven technische Artefakte direkt steuern.

[5] Siehe http://www.kevinwarwick.com

6 Künstliche Selbstwahrnehmung

Es wird zwischen *Top-Down-* und *Bottom-Up-Zugängen* zur künstlichen Intelligenz unterschieden [Dennett, 1998]. Der Top-Down-Zugang zu künstlicher Intelligenz ist es, intelligentes *Verhalten* zu programmieren. Dieser Zugang hat bisher nicht zu menschenähnlicher Kognition geführt, denn er endet mit dem *Frame-Problem*, das besagt, dass ein Programm genau das kann, was ihm der Programmierer an Verhalten einprogrammiert hat und darüber hinaus nicht anpassungsfähig ist. Selbst anpassungsfähige Systeme können sich nur an das anpassen, was ihnen programmiert wurde. Durch diese einprogrammierte Fehlleistung sind einer Top-Down-ausgerichteten Methode natürliche Grenzen diktiert. Der Bottom-Up-Zugang verwendet Evolution zur Entwicklung eines Systems. Man denke an ein Computerprogramm, das, entsprechend dem Kapitel über natürliche Selbstwahrnehmung, an einer willkürlichen Stelle der Evolution einsetzt und diese von dort simuliert. Durch diesen Prozess bekommt das Frame-Problem bekommt eine andere Form: Die Evolution ist abhängig von den Start- und Umgebungsbedingungen, so dass sich nur das evolviert, was für diese spezielle Simulationswelt *sinnvoll* ist. Es ist derzeit rechentechnisch unmöglich alle Facetten der Welt, die uns hervorgebracht hat, zu simulieren. Wir können also kein vollständiges Universum programmieren und schauen, ob Menschen dabei herauskommen. Während die Forschung der Artificial Intelligence zunächst mit dem Top-Down Zugang beschäftigt war, eröffnen die Rechenkapazitäten moderner Computer die verstärkte Nutzung von Bottom-Up oder hybriden Methoden.

6.1 Künstliche Metakognition

Was passiert, wenn wir das zuvor genannte Geistorgan der Buddhisten implementieren? Die hier vorgestellte Forschungsarbeit von Juan C. Zagal und Hod Lipson [Zagal and Lipson, 2009a][6] befasst sich damit, Agenten[7] zu schaffen, die ihre eigenen Gedanken als solche wahrnehmen und darüber reflektieren können. Hauptsächlich dient die Forschung dazu ei-

[6] Dieses Paper wurde bereits um ein weiteres, sehr ähnliches ergänzt, das sich hauptsächlich durch eine komplexere Umgebung (bewegte Lichtquellen) und ein komplexeres neuronales Netz, das selbstreferenzierende Knoten zulässt (*Recurrent Neural Network*), unterscheidet [Zagal and Lipson, 2009b]. Da diese zusätzliche Komplexität das Verständnis erschwert, jedoch im Rahmen meiner Arbeit nichts ergänzt, beschränke ich mich auf den simpleren Fall.

[7] Ein Agent ist eine Einheit, die autonom mit einer Umgebung interagiert. Ein Roboter, simuliert oder real, ist ein solcher Agent.

nem Agenten, dessen Steuerungseinheit unzugängig ist, ein neues Verhalten beizubringen, das besser an die Umwelt angepasst ist, ohne die alte Steuerungseinheit modifizieren zu müssen. Die Forscher simulierten einen Roboter, der aus einer Box mit einer freibeweglichen Kugel und zwei Rädern mit Motoren, sowie jeweils zwei Rotlichtsensoren und zwei Blaulichtsensoren bestückt ist. Der Roboter und eine Skizzierung des Versuchsablaufs sind in Abbildung 1 dargestellt. Dieser Roboter wird durch ein neuronales Netz gesteuert, das aus vier Eingangsknoten, die mit den Farbsensoren verknüpft sind, zwei Ausgangsknoten, verbunden mit je einem Motor und zwei Knoten als Hidden-Layer für zusätzliche Komplexität besteht. Abbildung 2 zeigt das Netzwerk mit Eingängen $z0 - z3$ und Ausgängen $u0, u1$. Dieses Netzwerk konstituiert den *Innate-Controller*, eine Reiz-Reaktion-Steuerungseinheit vergleichbar mit einem primitiven Nervensystem. Das „Konditionieren“ erfolgt, indem das Netzwerk als ein Genom mit 23 skalaren Parametern (14 Kantengewichte, 8 Schwellwerte und ein Motorskalierungsfaktor) repräsentiert wird und an einer künstlichen Evolution teilnimmt, von der das geeignetste Netzwerk nach 10.000 Generationen schließlich als Controller verwendet wird. Nach diesem Lernprozess hat der Roboter gelernt, rote Lichter zu meiden und verringert seine Geschwindigkeit, wenn er sich in der Nähe von blauen Lichtern befindet. Der Lernprozess dauert etwas länger als eine Stunde Simulation (extrapoliert aus anderen Zeitangaben).

Nun änderten die Wissenschaftler die Außenwelt, so dass das konditionierte Verhalten nicht mehr geeignet ist. Konkret tauschten sie die Bedeutung der Lichtquellen. Blaue Lichtquellen waren nun dringend zu meiden und rote zu suchen. Was gefordert wurde, ist also das gleiche Verhalten wie zuvor, nur bezogen auf umgekehrte Objekte. Das antrainierte Verhalten ist nicht länger zufriedenstellend und resultiert in einem Fitnessvergleichswert von 18,63. Um diesen Wert durch die gleiche Technik auf ca. 106 anzuheben, werden weitere ca. 4000-5000 Simulationen gebraucht, deren Zeit mit 30 Minuten angegeben ist. Wie die Wissenschaftler zeigten, kann dieses Niveau mittels der folgenden Technik in 40 Sekunden erreicht werden.

Ein weiteres neuronales Netzwerk, *Self-Model-Neural-Network* genannt und dargestellt auf Abbildung 3 wurde verwendet um den ersten Controller zu modellieren. Dieser Controller unterscheidet sich vom Innate-Controller dadurch, dass er mit einem erweiterten Hidden-Layer ausgestattet ist um mehr Komplexität zu ermöglichen. Das Training geschah ebenso über einen genetischen Algorithmus mit entsprechend grö-

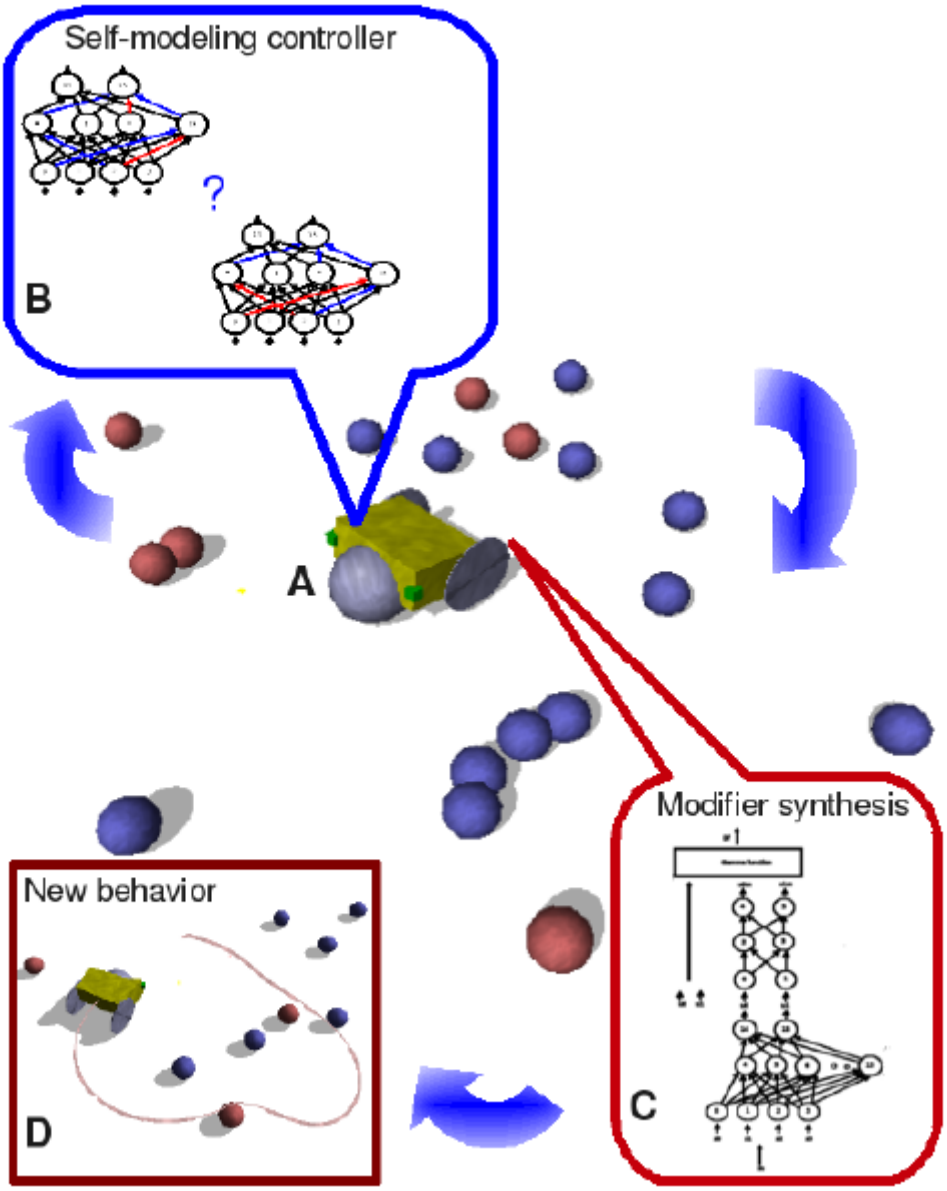

Abbildung 1: Der Roboter verhält sich gemäß seinem Innate-Controller (A), das Verhalten wird aufgezeichnet und zur Synthese eines Self-Model-Controllers verwendet (B), mit dessen Hilfe ein Modifier (C) erstellt wird, der in einem neuen Verhalten (D) resultiert. Quelle: [Zagal and Lipson, 2009a]

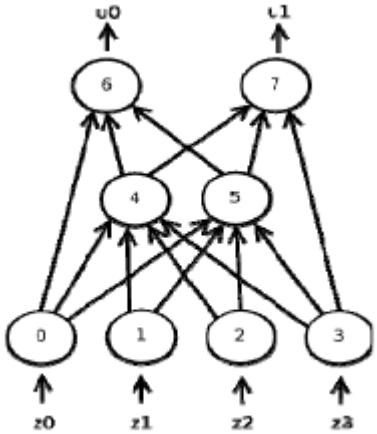

Abbildung 2: *Innate-Neural-Network*, das neuronale Netzwerk, das zur unmittelbaren Steuerung des Roboters verwendet wird. Quelle: [Zagal and Lipson, 2009a]

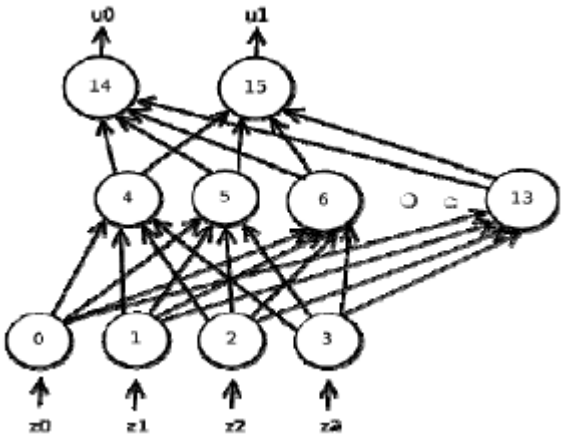

Abbildung 3: *Self-Model-Neural-Network*, das neuronale Netzwerk, das zur Modellierung des Innate-Controllers des Roboters verwendet wird. Quelle: [Zagal and Lipson, 2009a]

ßerem Genom. Ziel dieses Trainings war jedoch nicht das erfolgreiche Ausweichen bzw. Suchen von Lichtquellen, sondern dem ersten (jetzt unfähigen) Controller möglichst ähnlich zu werden. Gleiche Inputs sollten also gleiche Outputs erzeugen. Dazu ließen die Forscher den Roboter eine gewisse Zeit arbeiten und zeichneten Input/Output/Belohnungs Tripel auf, an die sie diesen Controller anpassten. Der Controller lernte also das alte (unangepasste) Verhalten.

Nun wurde ein drittes neuronales Netz, das *Modifier-Neural-Network* verwendet, um den Motor-Output des Innate-Controllers derartig zu modifizieren, dass die Belohnung steigt. Das verwendete Netzwerk ist in Abbildung 4 dargestellt, das Modifizierungssetup in Abbildung 5. Das Netz wird ebenso entwickelt wie die vorherigen beiden; als Input bekommt es alternativ den, zuvor aufgezeichneten, Output des Innate- oder des Self-Model-Netzes, wobei bei der Verwendung des Self-Model-Netzes zum Erreichen vergleichbarer Fitness ca. 25% Mehraufwand nötig ist, in dem die Entwicklung des Self-Model-Netzes selbst nicht einbezogen ist. Nach anderen Zeitangaben der Forscher macht dies etwa 10 Sekunden Unterschied aus. Der Extra-Aufwand für die Synthese des Self-Model-Netzes ist nicht angegeben, sollte sich aber in Grenzen halten, da keine realen bzw. simulierten Testläufe nötig sind, wie bei der Synthese des Innate-Controllers. Zur Modifizierung des Verhaltens wäre die Selbstmodellierung nicht nötig gewesen, sondern erhöht im Gegenteil den Aufwand im Vergleich zur direkten Synthese des Modifier-Netzes.

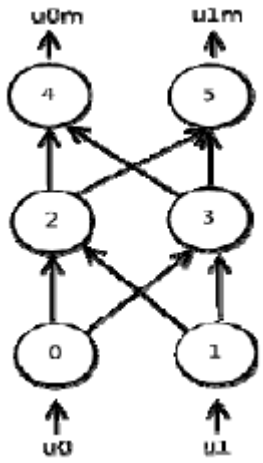

Abbildung 4: *Modifier-Neural-Network*, das neuronale Netzwerk, das zur Modifizierung des Verhaltens des Innate-Controllers verwendet wird. Quelle: [Zagal and Lipson, 2009a]

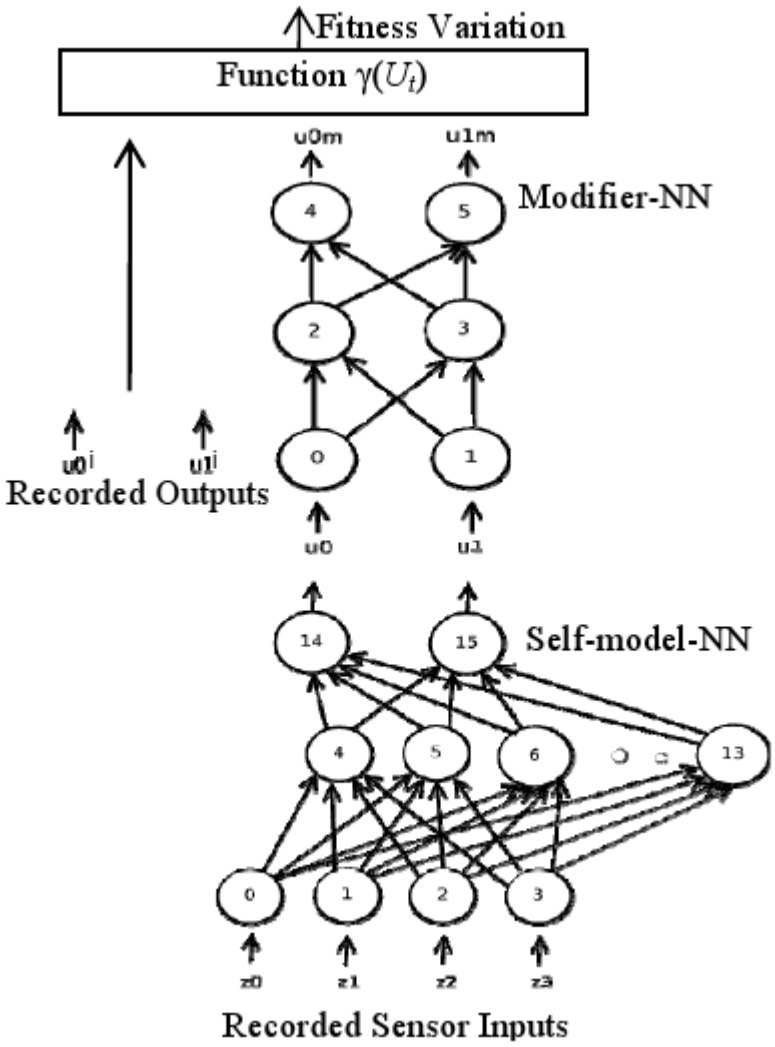

Abbildung 5: Der Aufbau der Zur Synthese des Modifier-Netzwerkes verwendet wird. Als Input dient wahlweise der aufgezeichnete Motor-Output des Innate-Controllers, oder des Self-Model-Controllers. Quelle: [Zagal and Lipson, 2009a]

Was diese Forschung demonstriert, ist, dass bereits primitive neuronale Netze in der Lage sind, Verhalten zu erlernen, ohne dieses Verhalten selbst erlebt zu haben. Ebenso ist kein kontinuierliches Lernen nötig, denn bereits aus diskreten Zeitintervallen war der Self-Model-Controller in der Lage das Verhalten zu Interpolieren. Diese Aspekte machen die Ergebnisse pragmatisch nutzbar für komplexe Systeme, die gewartet werden sollen, deren Funktionsweise allerdings nicht mehr bekannt ist. Es könnte sich als günstiger erweisen, das Verhalten von einem entsprechenden Controller erlernen zu lassen und das alte System dadurch zu ersetzen. Bewässerungsanlagen sind ein denkbares Beispiel. Was die Forschungsergebnisse leider nicht zeigen, ist warum die natürliche Evolution Metakognition für nötig empfand, denn offensichtlich kommt das Modifier-Netz sehr gut ohne Selbstmodellierung aus. Es wäre interessant zu beobachten, was geschieht, wenn der Innate-Controller wiederum vom Self-Model-Controller ein angepassteres Verhalten lernen könnte. Schlaf könnte für eine solche Rekonstruktion genutzt werden.

6.2 Künstliche Sozialevolution

Wie bereits bemerkt, ist Kooperation ein gern erforschtes Gebiet der Artificial-Life-Forschung. Die praktische Forschung reicht bis zum Austausch von Reflektionen. Es konnte gezeigt werden, dass sich Roboter, die über ein Modell ihrer Morphologie verfügen, schneller an Verletzungen anpassen können und dieses Wissen auch untereinander weitergeben können, wodurch die Anpassung erneut beschleunigt werden kann [Bongard, 2007]. Ein weiterer, vielversprechender, bisher jedoch scheinbar lediglich theorethisch verfolgter Ansatz, dessen Weg von Suzanne Sadedin und Greg Paperin in [Sadedin and Paperin, 2009] skizziert wurde und den ich in diesem Kapitel vorstellen möchte, ist die künstliche Sozialevolution. Zahlreiche Studien belegen (nach [Sadedin and Paperin, 2009]), dass große Gehirne sowohl mit menschenähnlicher Kognition, als auch mit speziellen sozialen Systemen zusammenhängen, allerdings nur schwach mit ökologischen Faktoren. Der Social-Brain-Theory [Dunbar, 1998] nach, entwickeln sich große Gehirne ausschließlich, wenn sie wirklich benötigt werden, da sie ein hohes Maß an Ressourcen benötigen. Es wird also ein starker Selektionsdruck benötigt, der ausgeprägte kognitive Fähigkeiten favoritisiert. Einen solchen Selektionsdruck sehen die Forscher in sozialen Systemen, die bestimmte Charakteristika aufweisen. Für die meisten Spezies sind hier körperliche Attribute ausreichend und selbst komple-

xe Schaustellung kann ohne komplexe Kognition erreicht werden. Wenn Macht jedoch über ein Netzwerk kooperierender Individialbeziehungen verteilt ist, so könne die Fähigkeit dieses Netzwerk zu manipulieren das Ziel der sozialen Selektion werden, was verbesserte Kognition zur Folge habe.

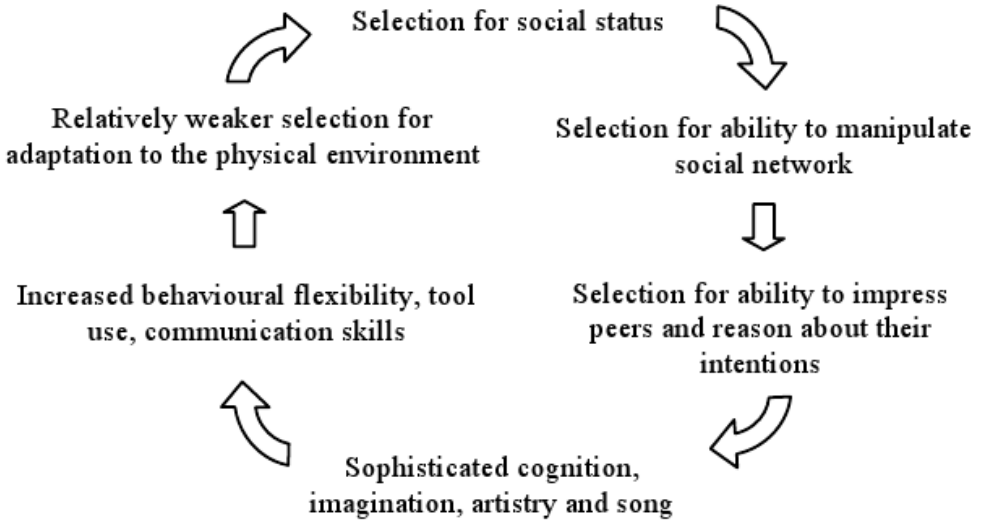

Abbildung 6: Konzeptuelles Modell der Evolution menschenähnlicher Kognition durch soziale Selektion. Quelle: [Sadedin and Paperin, 2009]

Abbildung 6 zeigt ein konzeptuelles Modell für die Feedbackschleife der Sozialevolution, die Sadedin und Paperin menschenähnlicher Kognition zugrunde liegen sehen. Beginnen wir oben im Kreis, bei der Selektion des sozialen Status, so bedeutet diese eine Selektion von sozialen Einheiten, die das soziale Netzwerk zu ihren Gunsten manipulieren können, was wiederum eine Selektion der Fähigkeit bedeutet, andere Knoten zu beeindrucken bzw. ihre Intentionen abzuschätzen. Diese Selektion führt zu verbesserter Kognition, Imagination, Kunst und Musik, sowie verbesserte Flexibilität im Verhalten, Werkzeugbenutzung und Kommunikationsfähigkeit. Diese führen wiederum zu einem schwächeren Selektionsdruck bezüglich Anpassung an die Außenwelt. Wir schließen den Kreis durch die, daraus verstärkte, Selektion auf Basis des sozialen Status. Als weiterer Katalysator sind den Forschern lange Beziehungen aufgefallen, die ein ausgeprägtes Balzverhalten mit sich bringen können. Dadurch werden sowohl der Selektionsdruck erhöht, als auch die kognitiven Fähigkeiten ausgeprägt.

Der Schritt zu einer künstlichen Sozialevolution ist jedoch kein einfacher. Die Forscher nennen vier Punkte für die Entwicklung menschenähnlicher Kognition:

- Ein komplexes Netzwerk sozialer Interaktion
- Selektion auf Basis des sozialen Status
- Fähigkeit mit anderen Mitgliedern des sozialen Netzwerkes zu kommunizieren
- Fähigkeit andere Mitglieder des sozialen Netzwerks und ihre Aktionen zu beobachten

Ausgehend von diesen Punkten führen die Wissenschaftler fort, dass künstliches leben - wenn bisher auch nicht intelligent - durch Koevulotion von Computerprogrammen[8] erzeugt werden kann und auch logische oder probablistische Systeme begrenzt repräsentatives „Denken“ demonstrieren und schlagen diese Forschungen als Kandidaten für Sozialevolution vor, da neuronale Netzwerke in erster Linie entstanden seien um Reaktionen auf Reize auszubilden und es nicht klar definierte Vorraussetzungen für weiterentwickelte Kognition geben könne, zu denen etwa assotiatives Gedächtnis oder einfache Formen der Empathie gehören können. Ich erachte den Vorschlag dieser Alternativen als wagemutig, da mir diese Vorraussetzungen in den genannten Alternativen, im Vergleich zu neuronalen Netzen, als deutlich weniger erfüllt erscheinen. Dennoch halte ich es für sinnvoll zu diesem Punkt der Forschung den Rahmen möglichst weit zu stecken und menschenähnliche Kognition nicht auf naturnahe bzw. naturimitierende Systeme zu beschränken. Die Forscher schließen mit Bemerkungen und Vorschlägen jeweils zur Fitness-Evaluation, Umgebung, Kommunikation und einem Ablaufplan, der die Evolution führt.

Die Herausforderung besteht in der Kreierung eines geeigneten Startpunktes. Die Forscher erwähnen im Bezug auf die Einheiten der Sozialevolution, dass unter Umständen gewisse Vorraussetzungen zu erfüllen seien, nicht jedoch aber konkret welche, oder wie sie diese erfüllen wollen. Hier sehe ich viel Spielraum für weitere Forschung.

[8] Ein freies (kostenlos und quelloffen) Programm dafür ist *Avida* (http://devolab.msu.edu/)

7 Schluss

Ich zeigte in dieser Arbeit, wie die natürliche Evolution von Selbstwahrnehmung, von ihrem Medium gelöst betrachtet und auf ein anderes Medium übertragen werden kann. Durch die hohe Rechenleistung moderner Computer wird der Bottom-Up Zugang - und damit die künstliche Evolution - eine realistische Unternehmung. Anhand zweier unterschiedlicher Arbeiten, im Forschungsgebiet Artificial Intelligence/Life, aus dem Jahr 2009, zeigte ich den aktuellen Stand der Forschung. Die Themen dieser Arbeiten scheinen im Vergleich zur Komplexität menschlicher Kognition bescheiden, sind in meinen Augen aber richtungsweisend für künftige interdisziplinäre Forschungsarbeit in Kollaboration von Natur- und Geisteswissenschaften, sei es als Cognitive Science oder Artificial Intelligence/Life. Evolution funktioniert und wird aktiv dafür verwendet, die Funktionsweise menschlicher Kognition zu erforschen. Diese Forschungsarbeit wirkt sich zwangsläufig auf unsere Kognition aus und kann uns Menschen in der Folge Erkenntnisse und Möglichkeiten schaffen, von denen wir derzeit tatsächlich nicht einmal zu träumen in der Lage sind.

Abbildungsverzeichnis

Literatur

[Andras, 2009] Andras, P. (2009). Networks of artificial social interactions. In *Proceedings of the 11th bi-annual European Conference on Artificial Life*.

[Bongard, 2007] Bongard, J. (2007). Exploiting multiple robots to accelerate self-modeling. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*.

[Dawkins, 2006] Dawkins, R. (2006). *The Selfish Gene*. Oxford University Press, 30th anniversary edition edition.

[Dennett, 1998] Dennett, D. C. (1998). *Brainchildren*. Penguin Books.

[Dunbar, 1998] Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190.

[Echterhoff, 2010] Echterhoff, M. (2010). Systemdenken, selbstorganisation und informationstechnologie? exemplarisches fallbeispiel: Bittorrent. Bachelor's Thesis, Universität Klagenfurt.

[Foote and Crystal, 2007] Foote, A. L. and Crystal, J. D. (2007). Metacognition in the rat. *Current Biology*, 17(6):551–555.

[Hanh, 1999] Hanh, T. N. (1999). *The Heart of the Buddha's Teaching*. Broadway Books.

[Humphrey, 1999] Humphrey, N. (1999). *A history of the Mind: Evolution and the birth of consciousness*. Copernicus.

[Iizuka et al., 2009] Iizuka, H., Ando, H., and Maeda, T. (2009). Cooperative behaviours with unknown partners by a homeostatic neural

controller. In *Proceedings of the 11th bi-annual European Conference on Artificial Life.*

[Kim and Lipson, 2009] Kim, K.-J. and Lipson, H. (2009). Towards a "theory of mindïn simulated robots. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation.*

[Maturana and Varela, 1987] Maturana, H. and Varela, F. (1987). *Der Baum der Erkenntnis.* Scherz Verlag.

[Phelps et al., 2009] Phelps, S., Nevarez, G., and Howes, A. (2009). The effect of group size and frequency-of-encounter on the evolution of cooperation. In *Proceedings of the 11th bi-annual European Conference on Artificial Life.*

[Sadedin and Paperin, 2009] Sadedin, S. and Paperin, G. (2009). Implications of the social brain hypothesis for evolving human-like cognition in digital organisms. In *Proceedings of the 11th bi-annual European Conference on Artificial Life.*

[Smith, 1986] Smith, B. C. (1986). Varieties of self-reference. In *Proceedings of the 1986 conference on Theoretical aspects of reasoning about knowledge.*

[Zagal and Lipson, 2009a] Zagal, J. C. and Lipson, H. (2009a). Self-reflection in evolutionary robotics: Resilient adaption wit a minimum of physical exploration. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation.*

[Zagal and Lipson, 2009b] Zagal, J. C. and Lipson, H. (2009b). Towards self-reflecting machines: Two-minds in one robot. In *Proceedings of the 11th bi-annual European Conference on Artificial Life.*

[Zaidel et al.,] Zaidel, E., Zaidel, D. W., and Bogen, J. E. The split brain. http://www.its.caltech.edu/ jbogen/text/ref130.htm Zugriff am 27.01.2010.