DAI-Labor
TU Berlin

Grundlagen der Künstlichen Intelligenz

08.12.2005: Modallogik

According to the intentional stance, an agent is assumed to decide to act and communicate based on its beliefs about its environment and its desires and intentions.
nach D.C. Dennet: The Intentional Stance

Dr.-Ing. Stefan Fricke
stefan.fricke@dai-labor.de

AIOT
Agententechnologien in betrieblichen Anwendungen und der Telekommunikation

Lernziele:

Gliederung

- ⇨ **Einführung in Modallogik**
- ⇨ **Modallogiken für Zeit und Handeln**
- ⇨ **Intentionalität**
- ⇨ **Epistemische Logik mit Belief, Desire, Intention**
- ⇨ **BDI-Agenten**
- ⇨ **Zusammenfassung**

Probleme klassischer Logik

Klassische Logik beschreibt Tatsachen und Zusammenhänge, aber..

- ⇨ **… kein Handeln, keine Prozesse**
  - → Der Effekt einer Aktion ist unvorhersagbar in nichtdeterministischen und in dynamischen Umgebungen
- ⇨ **… keine Zeit**
- ⇨ **… keine Überzeugungen, Motivationen, etc**
  - → Ich hoffe, du glaubst, dass ich weiß, dass der Zahnarztbesuch Schmerzen bereitet. Ich beabsichtige nicht, Schmerzen zugefügt zu bekommen. (Aber ich beabsichtige zum Zahnarzt zu gehen)

etwas tun wird im Situationskalkül durch Vor- und Nachbedingungen als Zustand beschrieben. Der Effekt einer Aktion ist nicht immer vorhersehbar, beispielsweise kann eine Aktion scheitern oder kann das Ergebnis eines Würfelwurfs nicht vorhergesagt werden.

… keine Zeit: Wie lange brauche ich zum Bahnhof?

… keine Überzeugungen, Motivationen, etc: Ich hoffe, du glaubst, dass es morgen regnen wird. Ich glaube, du weißt, dass ich weiss, dass …

Modallogik — Einführung

- ⇨ **Idee: Aus logischen Zusammenhängen und Inferenzen der Prädikatenlogik sinnvolle Sachverhalte und Folgerungen durch Einführung so genannter Modaloperatoren ermöglichen.**
- ⇨ **Modaloperatoren werden vor eine Formel geschrieben und geben ihr eine neue Interpretation. Z. B. für p = es_regnet**

ICH_GLAUBE p vs. ICH_HOFFE ¬p

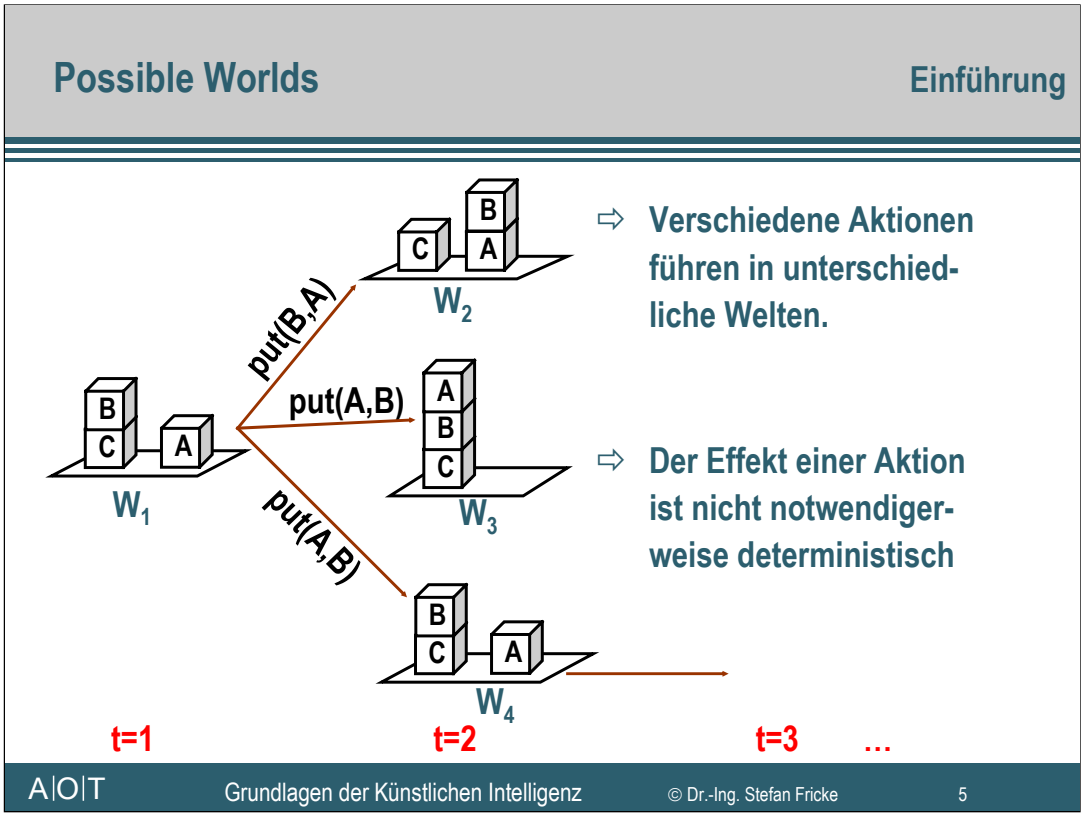
IRGENDWANN p und ICH_WEISS IRGENDWANN p

ICH_GLAUBE DU_HOFFST p vs. ICH_HOFFE DU_GLAUBST p

Die Worte in GROSSBUCHSTABEN sind Modaloperatoren. Offensichtlich ergibt sich eine andere Bedeutung als die des Wissens von p: IRGENDWANN drückt Zeit aus (eine zukünftige Possible World), ICH_HOFFE einen Wunsch, ICH_GLAUBE einen Belief, ICH_WEISS Wissen.

Modallogiken erweitern Prädikatenlogik um Modal-Operatoren. Sie erlauben sinnvolle Schlussfolgerungen und insbesondere auch nichtmonotones Schließen.

Epistemische Logiken nutzen Belief-Operator (*Wissen* über die Welt). selbstbezüglicher *Glauben* bzgl. des eigenen Wissens. Agent kann Schlüsse über die Welt mit dem Bewusstsein eigener Informationsdefizite ziehen. Ein Agent glaubt ein Fakt genau dann, wenn es in seiner Belief-Datenbasis enthalten ist (Vollständigkeitsannahme)



Ausgehend vom aktuellen Zustand beschreiben die Possible Worlds einen Baum. Dessen Zweige stellen alternative „Lebenspfade“ dar.

Verschiedene Aktionen führen in unterschiedliche Welten: put(B,A) führt in die Welt W2, put(A,B) in die andere Welt W3.

Der Effekt einer Aktion ist nicht notwendigerweise deterministisch: z.B. kann das put(A,B) mit einer gewissen Wahrscheinlichkeit scheitern (wenn der Roboterarm die Aktion nicht korrekt durchführt). Ergebnis ist eine andere (hier unveränderte) Welt W4.

- **Erreichbarkeitsrelation $R(W_1, W_2)$: W_2 ist von W_1 aus erreichbar.**
- **Möglichkeit: $\Diamond P$ ist folgerbar in einer Welt W genau dann, wenn P wahr ist in *mindestens* einer möglichen Welt**
  - $\exists w': R(w,w') \wedge w' \models P$
- **Notwendigkeit: $\Box P$ ist folgerbar in einer Welt W genau dann, wenn P wahr ist in *jeder* möglichen Welt**
  - $\forall w': R(w,w') \Rightarrow w' \models P$

Quelle für Beispiele: http://www.informatik.uni-leipzig.de/~duc/Thesis/node9.html

Axiomsschemata für Modallogiken

K:	$\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$	(Distribution)
D:	$\Box A \rightarrow \neg\Box\neg A$	(Konsistenz)
T:	$\Box A \rightarrow A$	(Reflexivität)
B:	$A \rightarrow \Box\Diamond A$	(Symmetrie)
4:	$\Box A \rightarrow \Box\Box A$	(Transitivität)
NEC:	$A \rightarrow \Box A$	(Notwendigkeit)

Es gelten ferner alle Tautologien der Aussagenlogik.

D,T,B,4,NEC, sowie weitere, hier nicht aufgeführte Axiome stellen Einschränkungen der durch K axiomatisierten Modallogik dar. K gilt in jeder Modallogik und stellt damit quasi die Mindestanforderung dar.

D ist schwächer als T. T drückt Wissen aus und klappt nicht mit Glauben.

K: If you believe that p implies q then if you believe p then you believe q. K ist das kleinste Modallogische System.

Andere Def für D mit Belief: $\Box A \rightarrow \neg\Box \neg A$ (if you believe p the you do not believe that p is false.)

D ist „seriell“

Possible Worlds Semantik

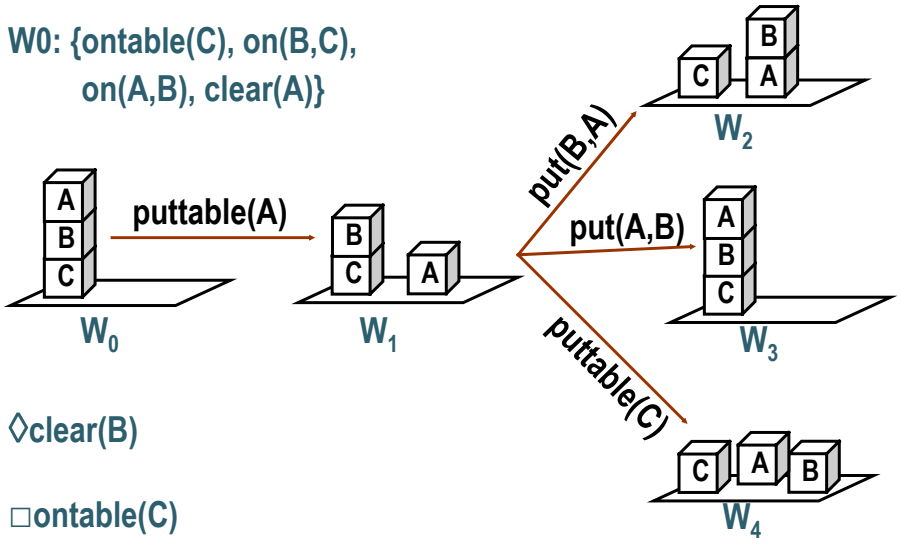
- **Ein Modell ist nicht nur eine, sondern eine Menge von Welten …**
  - … generiert durch Aktionsauswahl,
  - … generiert durch eine Sequenz von Aktionen,
  - … generiert durch mehrere mögliche Ergebnisse einer Aktion,
  - … generiert durch Wissenslücken.
- **Possible Worlds dienen zur Beschreibung dieses Modells…**

Wissenslücken: siehe Default Logic vom letzten Mal.

Jede Interpretation der klassischen Logik wird nun zur Interpretation einer möglichen Welt.

Beispiel für Possible Worlds in der Blockswelt

⇨ W0: {ontable(C), on(B,C), on(A,B), clear(A)}



⇨ ◊clear(B)

⇨ □ontable(C)

$W0 = \{on(A,B), on(B,C), clear(A), ontable(C)\}$

$W1 = \{ontable(A), clear(B)\ on(B,C), clear(A), ontable(C)\}$

$W3 = \{ontable(A), clear(B)\ on(B,C), clear(A), ontable(C)\}$

$W3 = W0$

$W4 = \{ontable(A), ontable(B), ontable(C), clear(A)\ clear(B)\ clear(C)\}$

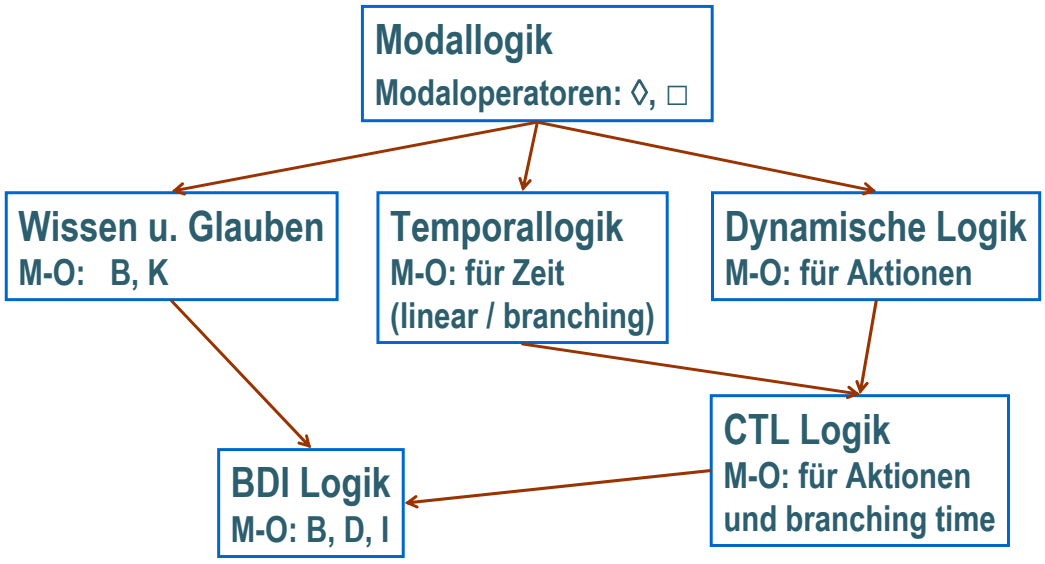
◊clear(B) gilt in W1, W2 und W4 nicht abe in W0 und W3

□ontable(C) gilt tatsächlich in allen 5 Welten.

Gliederung

- ⇨ **Einführung in Modallogik**
- ⇨ **Modallogiken für Zeit und Handeln**
  - → Temporallogik
  - → Dynamische Logik
- ⇨ **Intentionalität**
- ⇨ **Epistemische Logik mit Belief, Desire, Intention**
- ⇨ **BDI-Agenten**
- ⇨ **Zusammenfassung**

Modallogiken im Zusammenhang



Logik von Glauben und Wissen

- $K_a\varphi$ Agent a weiß φ
- $Bel_a\varphi$ Agent a glaubt φ

- Ein Agent weiß / glaubt eine Proposition φ genau dann, wenn φ in allen möglichen Welten gilt.

Eigenschaften von Wissen und von Glauben

(1) $K_a\varphi \wedge K_a(\varphi \rightarrow \psi) \rightarrow K_a\psi$

(2) $K_a\varphi \rightarrow \varphi$ — **Wissen**

(3) $K_a\varphi \rightarrow K_a(K_a\varphi)$ — **positive Introspektion**

(4) $\neg K_a\varphi \rightarrow K_a(\neg K_a\varphi)$ — **negative Introspektion**

⇨ **Für $Bel_a\varphi$ analog:**

- gilt (1)
- gilt nicht (2)
- gilt (3)
- kann (4) gelten, ist aber problematisch.

(4) drückt Allwissenheit aus

Beispiel

- 2 Personen sitzen einander gegenüber und wissen (1), dass mindestens einer von ihnen eine weiße Stirn hat (2). B sagt, er wisse nicht, ob er eine weiße Stirn habe (3). Daraus folgert A, dass er selbst eine weiße Stirn hat.

1. $K_A(\neg White(A) \rightarrow K_B(\neg White(A))$ (1)
2. $K_A(K_B(White(A) \vee White(B)))$ (2)
3. $K_A(\neg K_B(White(B)))$ (3)

- Kann A daraus schließen, dass er selbst eine weiße Stirn hat?

Das two wise men Problem nach Genesereth und Nilsson

(1): A weiß: Wenn A keine weiße Stirn hat, dann sieht dies B und weiß es entsprechend.

(2): A weiß, dass B weiß, dass mindestens einer der beiden eine weiße Stirn hat.

(3): A weiß, dass B die Farbe seiner Stirn nicht weiß (da er sie nicht sehen kann).

Beispiel

Inferenzregeln **A1**: $K_A\varphi \wedge K_A(\varphi \rightarrow \psi) \rightarrow K_A\psi$; **A2**: $K_A\varphi \rightarrow \varphi$
R2: $\varphi \rightarrow \psi \wedge K_A\varphi \rightarrow K_A\psi$

1. $K_A(\neg White(A) \rightarrow K_B(\neg White(A))$	
2. $K_A(K_B(White(A) \vee White(B)))$	
3. $K_A(\neg K_B(White(B)))$	
4. $\neg White(A) \rightarrow K_B(\neg White(A))$	1, A2; $X \rightarrow Y \Leftrightarrow \neg X \vee Y$
5. $K_B(\neg White(A) \rightarrow White(B))$	2, A2
6. $K_B(\neg White(A)) \rightarrow K_B(White(B))$	5, A1
7. $\neg White(A) \rightarrow K_B(White(B))$	4, 6
8. $\neg K_B(White(B)) \rightarrow White(A)$	Umkehrschluss von 7
9. $K_A(White(A))$	3, 8, R2

R2 ist problematisch, weil es Allwissenheit ausdrückt.

1.-3. ist die Wissensbasis der vorherigen Folie (das Ausgangswissen).

4. Etwas wissen heißt, dass es wahr ist (K-Modaloperator entfernen)

5. K-Modaloperator für B entfernen und umformen

6. ist keine direkte Resolvente: 5. ist die Implikation, und zusammen mit der Prämisse (¬White(A)) ist 6. korrekt.

7. Die Implikation von 4 ist die Prämisse von 6.

8. A->B <=> ¬B->¬A

9. Anwendung der “problematischen” Inferenzregel R2: Aus wahren Sachverhalten entsprechendes Wissen ableiten (Omniscience).

Dynamische Logik

- ⇨ **Aktionssymbole a, b, …**
- ⇨ **a;b** — **Sequenz**
- ⇨ **a+b** — **nichtdeterministische Auswahl**
- ⇨ **p?** — **deterministische Auswahl (IF-THEN)**
  Aktion, basierend auf Wahrheitswert von p
- ⇨ **a*** — **0 oder mehr Wiederholungen von a**
- ⇨ **<a>p** — **a macht p möglicherweise wahr (analog zu ◊)**
- ⇨ **[a]p** — **a macht p notwendigerweise wahr (analog zu □)**

Aktionslogik

nichtdeterministische Auswahl = exklusiv ODER (entweder a oder b)

Es gilt: $\langle A\rangle P \leftrightarrow \exists w'\colon R_A(w,w') \wedge P$ entailed in w'

und: $[A]P \leftrightarrow \forall w'\colon R_A(w,w') \rightarrow P$ entailed in w'

$M_4 \models_{s,t} x[a]p$ iff $(\forall t' \in s\colon [s;t,t'] \in |a|^x \Rightarrow M_4 \models_{s,t'} p)$ ***p*** is true on all the set of moments ***t'*** on a given path ***s*** starting at the current moment ***t*** while agent ***x*** executes action ***a***

$M_4 \models_{s,t} x\langle a\rangle p$ iff $(\exists t' \in s\colon [s;t,t'] \in |a|^x \,\&\, M_4 \models_{s,t'} p)$ ***p*** is true at a moment ***t'*** on a given path ***s*** starting at the current moment ***t*** while agent ***x*** executes action ***a***

Temporallogik

- ⇨ **Momente T mit einer partiellen Ordnung <, die die Vorher-Beziehung ausdrückt.**
  - → Jeder Moment entspricht einer Possible World
- ⇨ **p U q** — **p ist wahr bis q wahr wird (UNTIL)**
- ⇨ **Xp** — **p ist im nächsten Moment wahr (NEXT)**
- ⇨ **Pp** — **p war in einem früheren Moment wahr (PAST)**
- ⇨ **Ep** — **p ist irgendwann in der Zukunft wahr (EVENTUALLY)**
- ⇨ **Ap** — **p wird immer wahr sein (ALWAYS)** **$Ap \equiv \neg E \neg p$**

Ap: in allen möglichen Welten ab der Gegenwart ist p wahr.

Man unterscheidet zwischen linearer und verzweigender Logik. Lineare Temporallogik geht von Sequenzen von Events bzw. Outcomes aus, während verzweigende Logik verschiedene aufspannende Sequenzen in Form possible Worlds betrachtet (also Bäume).

$M_4 \models_{s,t} \mathbf{A}\, p$ iff $(\forall s: \; s \in S_t \Rightarrow M_4 \models_{s,t} p)$ **s** is a path, $\mathbf{S_t}$ - all paths starting at the present moment

Gliederung

- ⇨ Einführung in Modallogik
- ⇨ Modallogiken für Zeit und Handeln
- ⇨ **Intentionalität**
- ⇨ Epistemische Logik mit Belief, Desire, Intention
- ⇨ BDI-Agenten
- ⇨ Zusammenfassung

Fragestellungen zum Verhalten komplexer Systeme

- **Wie lässt sich das Verhalten eines Systems vorhersagen ohne Kenntnis seiner internen Struktur?**
- **Welche Repräsentationen eignen sich zur Definition**
  - des beobachtbaren Verhaltens?
  - des erwarteten Verhaltens?
- **Die Intentionalität gibt Antworten auf die Beschreibung des Verhaltens komplexer Systeme...**

Der Agent als Black-Box: Ein mit Sensoren (Inputs) und Aktoren (Outputs) versehenes System, das ein anderer programmiert hat. Die erste Frage ist demnach sinnvoll für alle komplexen Systeme. Es geht darum, einerseits Erklärungen zu finden, warum ein Agent in der Vergangenheit so gehandelt hat wie es beobachtet wurde und vor allem auch darum, in die Zukunft gerichtete Aussagen zu treffen. Diese Fragestellungen führen uns direkt zum Thema Intentionalität...

Zur Theorie des Intentional Stance

⇨ **Der Philosoph Daniel Dennett unterscheidet zwischen drei Heransgehensweisen, wenn man Vorhersagen über ein System treffen will:**

- → **physical stance,**
- → **design stance,**
- → **intentional stance...**

stance = Einstellung, Haltung

Die 3 Sichten eignen sich sowohl für Vorhersagen, als auch für Verhaltensbeschreibungen und Verhaltenserklärungen.

Physical Stance

- Der **physical stance** beschreibt Entitäten in physikalischen Begriffen.
- Zum Beispiel
  geometrische Körper in der Mathematik,
  Mechanik im Maschinenbau,
  Statik in der Architektur.

Objekte mit eindeutigen physikalischen Eigenschaften lassen sich über den physical stance beschreiben. Eine derartige Beschreibung hat den Vorteil, exakt zu sein.

physical stance: "simply the standard laborious method of the physical sciences" (Dennett, 1996, S. 28);

Design Stance

- ⇨ Der **design stance** nimmt an, dass ein Objekt zu einem Zweck entwickelt wurde und sich entsprechend verhält.
- ⇨ Die Erklärung erfolgt anhand funktionaler Begriffe.
- ⇨ Beispiel: Das Verhalten eines Autos wird über Funktionen wie Gas geben, Kupplung treten, Lenken, Bremsen beschrieben.

Die Design-Einstellung beschäftigt sich auch mit der Frage, wie ein System aus einzelnen Komponenten zusammengesetzt ist (mechanistische Sicht).

design stance: "that an entity is designed as I suppose it to be, and that it will operate according to that design" (Dennett, 1996, S. 29).

Intentional Stance

- **Der intentional stance beschreibt das Verhalten einer Entität als das eines rationalen Agenten.**
  - Es ist immer auch möglich, dasselbe Verhalten in rein physikalischen oder funktionalen Begriffen zu erklären.
- **Aus pragmatischen Gründen ist die intentionale Perspektive jedoch unverzichtbar. Was können das für pragmatische Gründe sein?**
- **Die Komplexität der Entität.**
  - U.U. ist es sinnvoll, einem Thermostaten die „Absicht“ zuzubilligen, die Temperatur konstant zu halten.

intentional stance: "as if it were a rational agent" (Dennett, 1996, S. 31).

In der Umgangssprache finden intentionale Zuschreibungen häufig statt, wenn Maschinen nicht funktionieren (z.B. „der Motor will nicht starten“).

Das Beispiel mit dem Thermostaten geht auf John McCarthy zurück, einem Pionier der Informatik (z.B. Erfinder der Programmiersprache LISP):

Dieser Thermostat, dessen Funktion es ist, die Temperatur eines Raumes zu regulieren, hätte also drei Möglichkeiten, etwas zu *glauben*:

1. „Der Raum ist zu kalt.“,
2. „Der Raum ist zu warm.“,
3. „Der Raum ist genau richtig temperiert.“.

Außerdem hätte es einen einzigen *Wunsch*, nämlich die Temperatur im Raum der an ihm eingestellten anzupassen. Der Clou bei dieser ungewöhnlichen Betrachtungsweise ist nicht zuletzt der, dass man durch diese Beschreibung alle möglichen Thermostate adäquat beschreiben kann, während, man mit der physikalischen oder funktionalen Einstellung immer nur ganz spezielle Implementationen fokussiert.

Weitere Beispiele für Intentionalität: Ein Rehkitz ist durstig und *hat die Absicht* an die Zitze der Mutter zu gelangen.

His *belief* that the gun was loaded caused his heart attack

Theorie des Intentional Stance

- **Die Kernidee ist, das Verhalten von Individuen durch die Zuschreibung von Wünschen und Überzeugungen rational verständlich und damit prognostizierbar zu machen.**
- **Diese Zuschreibungen beschreiben jedoch keine Tatsachen, aus denen das Verhalten aufgrund kausaler Gesetzmäßigkeiten erschlossen werden kann.**
  - Vielmehr stellen intentionale Beschreibungen ein Abstraktionswerkzeug dar, das mit familiären Begriffen operiert.

Unter Intentionalität versteht man in der Philosophie eine besondere Eigenschaft von Überzeugungen, Wünschen und anderen psychischen Vorkommnissen.

Überzeugungen und Wünsche müssen nicht in jedem Fall erfüllbar sein bzw. eintreten. D.h., mit klassischer Logik kommt man hier nicht weiter.

Man spricht häufig auch vom „mentalen Zustand“ eines Agenten. Gemeint ist seine „geistige Verfassung“ (kennen gelernt haben wir in vorangegangenen Vorlesungen schon die mentalen Begriffe „Ziele“, „Beliefs“ und „Intentionen“) und nicht etwa der gesamte Programmzustand, wie er in Form eines Hexdumps vorliegen könnte.

Anders als in der klassischen Logik werden an intentionale Begriffe schwächere Kriterien hinsichtlich Konsistenz, Erfüllbarkeit usw. angelegt.

Schachcomputer als Beispiel für den Intentional Stance

- **Wie setzt man sich adäquat mit einem Schachcomputer auseinander, dessen exakte innere Implementation man nicht kennt?**
  - Er hat *Wissen* über die Figuren auf den Feldern und die Bewegungsmöglichkeiten der Figuren;
  - Er hat den *Wunsch*, das Spiel zu gewinnen.
- **Mit diesen Zuschreibungen kann man gute Vorhersagen oder Erklärungsansätze über sein Verhalten liefern.**

Bei sehr komplexen und immer intelligenteren Systemen bietet diese Haltung die einzige praktikable Möglichkeit sinnvolle Charakterisierungen zu treffen.

Gliederung

- ⇨ **Einführung in Modallogik**
- ⇨ **Modallogiken für Zeit und Handeln**
- ⇨ **Intentionalität**
- ⇨ **Epistemische Logik mit Belief, Desire, Intention**
  - → Modaloperatoren Belief, Desire, Intention
  - → Axiomatisierungen für B, D, I
- ⇨ **BDI-Agenten**
- ⇨ **Zusammenfassung**

Belief, Desire, Intention (BDI) bietet dreierlei:

- ⇨ **ein philosophisches Modell des menschlichen Denkens und Handelns,**
  - → [Bratman, 1987]
- ⇨ **verschiedene Architekturimplementierungen,**
  - → (IRMA, PRS, JACK, JIAC)
- ⇨ **eine abstrakte logische Semantik.**

Mit dem philosophischen Modell, das als Grundlage für Architekturen und Semantik zitiert und herangezogen wird, beschäftigen wir uns nicht. Verschiedene BDI-Architekturen werden später vorgestellt (JIAC ist eine ebensolche). Zunächst aber der theoretische Hintergrund, die Semantik hinter BDI...

[Bratman, 1987]: Michael E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

Epistemische Modallogiken für Agenten

Epistemische Logik nutzen, um

- ⇨ **mentale Attitüden von Agenten zu modellieren:**
  - → Beliefs, Desires, Goals, Know How, Intentions, …,
- ⇨ **das Verhalten eines Agenten zu begründen,**
- ⇨ **das Verhalten eines Agenten vorhersehbar zu machen,**
- ⇨ **sinnvolle Handlungen zu generieren.**

Limited rationality / limited computational resources means that the agent can't derive everything that it "believes"

B,D,G beziehen sich auf Welten; I auf Pfade.

Belief, Desire, Intention

- **$Bel_x p$: Agent x glaubt p.**
- **$Des_x p$: Agent x wünscht Zustand p**
  - Ziele ermöglichen in die Zukunft gerichtetes Handeln.
- **$Int_x p$ Agent x beabsichtigt, den Zustand p zu erreichen**
  - Intentionen sind mit Handlungen verknüpft

Beliefs drücken aus, was ein Agent über die Welt weiß. In einer dynamischen Welt haben Aussagen begrenzte Gültigkeit: Was ein Agent eben noch geglaubt hat, muss im nächsten Zeitpunkt nicht mehr wahr sein, d.h., die Monotonie-Annahme gilt nicht. Eine lokale Sicht auf die Welt ist notwendigerweise unvollständig. Was ein Agent nicht weiß ist nicht unbedingt unwahr, sondern meistens nur unbekannt.

Aus $B_x p$ folgt nicht automatisch $B_x B_x p$

Zumeist (und im Folgenden) werden Desires und Goals synonym verwendet.

Ziele ermöglichen in die Zukunft gerichtetes Handeln: Ein Ziel zu haben ist die Antriebskraft (genauer: Lenkungskraft), die einem hilft, vorwärtszukommen (und zwar zielgerichtet). Ohne Ziele treibt man umher, mit Zielen kann man auch Auskunft darüber geben, wie weit man von seinen Vorstellungen noch entfernt ist und welche Handlungen zu planen sind, um das Ziel zu erreichen. Ziele können auch zum Zwecke der gegenseitigen Koordination kommuniziert werden – sie stellen dabei eine geeignete Abstraktion von der konkreten Welt dar (die vollständig zu beschreiben zu aufwändig wäre).

Eigenschaften von Zielen (Goals, bzw. Desires)

- ⇨ **Desires können nicht erfüllbar oder gar inkonsistent sein.**
  - → Goals sind Teilmengen von Desires: erreichbar und konsistent.
- ⇨ **Der Agent soll glauben, dass sein Ziel erreichbar ist.**
  - → Das verhindert, dass der Agent Ziele annimmt, von denen er glaubt, dass sie unerreichbar sind.
  - → Diese Eigenschaft wird Realismus genannt.
- ⇨ **Ziele (und auch Intentionen) können überprüft werden.**

Widersprüchliche Zustände sind wünschbar und der Agent muss nicht wissen, wie er p erreicht.

Beispiel für inkonsistente Ziele: noch diese Woche ins Theater gehen wollen und noch diese Woche in die Oper gehen wollen – inkonsistent, wenn die Woche nur noch einen Tag hat.

In der Realität sind Wünsche und Ziele verschiedene Begriffe. Über den Zusammenhang herrscht Uneinigkeit:

-Wünsche haben einen weniger rationalen und zu Handlungen verpflichtenden Charakter als Ziele.

-Wünsche können im Gegensatz zu Zielen auch inkonsistent sein.

-Ein Wunsch ist ein Ziel, das durch eine Aktion erfüllt wird.

-Singh analysiert, dass Wünsche einen weniger rationalen und zu Handlungen verpflichtenden Charakter haben als Absichten.

-Wünsche können im Gegensatz zu Intentionen (und Beliefs) auch inkonsistent sein.

-Cohen & Levesque definieren einen Wunsch als ein Ziel, das durch eine Aktion erfüllt wird.

-Rao & Georgeff behandeln Wünsche und Ziele scheinbar synonym, bevorzugen aber den Begriff´‚Ziel'.

-Singh beschreibt eine Verwandtschaft von Zielen mit Wünschen, Strategien und Intentionen, ohne jedoch näher darauf einzugehen. Aus begrifflichen und methodischen Gründen beschränkt er sich auf die Konzepte von Glauben und Absichten.

-Rao & Georgeff formulieren eine genaue Abhängigkeit zwischen den drei Schichten der Glaubens-, Wunsch- und Ansichtswelt. (s.u.)

Dadurch, dass ein Ziel explizit (und nicht etwa implizit wie bei einem Thermostaten) repräsentiert wird, kann der Agent testen, ob das Ziel bereits erreicht wird, welche Aktion ihm dem Ziel am nächsten bringt oder ob Zielkonflikte vorliegen.

Eigenschaften von Intentionen

- ⇨ **Jede Intention soll erfüllbar sein.**
- ⇨ **Die Intentionen sollen gegenseitig konsistent sein.**
- ⇨ **Sie sollen mit den Beliefs des Agenten vereinbar sein.**
  - → $Int_x p$ impliziert nicht $Des_x p$
- ⇨ **Ein Agent muss nicht unbedingt an die Erfüllbarkeit einer Intention glauben, ausschließen darf er sie jedoch nicht.**

Eigenschaften nach Singh/ Rao/ Georgeff.

Im Gegensatz zu Zielen ist es eine Anforderung an Intentionen, gegenseitig konsistent sein. Der Grund ist einfach, Intentionen sind die beabsichtigten Handlungen des Agenten, und diese Handlungen sollten sich nicht behindern. Bei Zielen ist die Lage anders, der Agent hat zwar widersprüchliche Absichten, aber noch keine konkreten Handlungen zum Erreichen aller dieser Absichten geplant. Er kann bestimmte Ziele einfach zurückstellen und später verfolgen.

Intentionen sollen mit den Beliefs des Agenten vereinbar sein (nicht notwendigerweise mit den Zielen): Singh benutzt zur Veranschaulichung der Abgeschlossenheit das Beispiel eines Bomberpiloten, der *beabsichtigt*, eine Munitionsfabrik zu zerstören und gleichzeitig *weiß*, dass dabei eine benachbarte Schule unweigerlich mit zerstört würde, deren Zerstörung er keineswegs beabsichtigt. Eine solche Konstellation soll erlaubt werden.

Intentionen sind persistent. Sie lenken die Handlungsplanung und beschränken das notwendigt Reasoning (es muss nicht immer alles neu überdacht werden).

Der Agent muss hinsichtlich seiner Pläne und Ziele verpflichtet sein, wenn er sie annimmt. Er muss außerdem in der Lage sein, Pläne und Ziele zu überdenken (in bestimmten Entscheidungssituationen unmögliche Intentionen fallen lassen, nicht erreichte Intentionen aktiv lassen). Diese „committeten“ Ziele und Pläne sind die Intentionen des Agenten.

Possible Worlds in der Computation Tree Logic (CTL)

- ⇨ Eindeutige Vergangenheit, verzweigende Zukunft
- ⇨ Ein **Pfad** ist eine maximale Menge von Momenten, ausgehend von der Gegenwart bis in alle Momente in der Zukunft, entlang einem Zweig gemäß <
- ⇨ **Situation** = Welt zu einem Zeitpunkt
- ⇨ **Zustandsformeln** beziehen sich auf eine Situation
- ⇨ **Pfadformeln** beziehen sich auf einen Pfad …

Zustandsformeln: U, X, P und E aus der Temporallogik

Pfadformeln: A und E aus der Temporallogik. Es gilt: $Ep \equiv \neg A \neg p$

Operatoren für Possible Worlds

- **Modale Operatoren für Situationen:**
  - **A** *a* (Always) In jeder möglichen Zukunft gilt *a*
  - **E** *a* (Eventually) In einer possible world gilt *a*
  - z.B. E B p (irgendwann glauben, dass es regnet)

- **Modale Operatoren für Pfade:**
  - **O** *f* (optional) für mindestens einen Pfad ist *f* wahr
  - **I** *f* (inevitable) für alle ausgehenden Pfade ist *f* wahr
  - z.B. O A B p (in einem möglichen Leben immer glauben, dass es regnet)

I/O beziehen sich auf Pfade; A/E auf Situationen.

A = in allen zukünftigen Pfaden (inevitable)

E = in mindestens einem zukünftigen Pfad (eventually)

$M_4 \models_{s,t} x[a]p$ iff $(\forall t' \in s: [s;t,t'] \in |a|^x \Rightarrow M_4 \models_{s,t'} p)$ ***p*** is true on all the set of moments ***t'*** on a given path ***s*** starting at the current moment ***t*** while agent ***x*** executes action ***a***

$M_4 \models_{s,t} x<a>p$ iff $(\exists t' \in s: [s;t,t'] \in |a|^x \ \& \ M_4 \models_{s,t'} p)$ ***p*** is true at a moment ***t'*** on a given path ***s*** starting at the current moment ***t*** while agent ***x*** executes action ***a***

$M_4 \models_{s,t}$ **A** p iff $(\forall s: \ s \in S_t \Rightarrow M_4 \models_{s,t} p)$ **s** is a path, $\mathbf{S_t}$ - all paths starting at the present moment

$M_4 \models_{s,t} (V\ a : p)$ iff $(\exists x: x \in Ag$ and $M_4 \models_t p|^a_x)$ there is an agent, be it ***x***, capable of executing ***a*** under which ***p*** comes true, if ***a*** is executed at ***t***

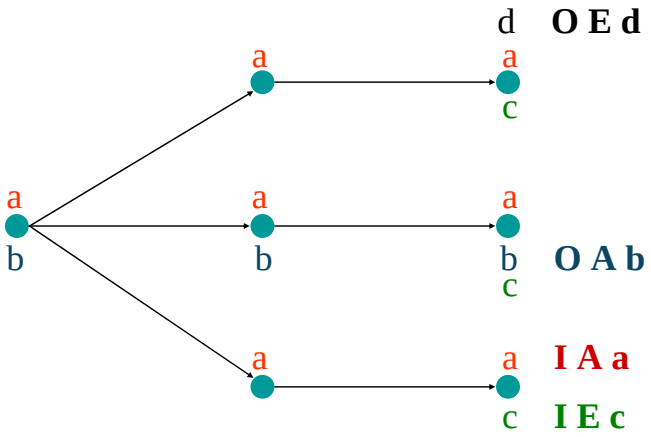
optional entspricht dem Existenz-Quantor, inevitable dem All-Quantor.

Weitere, hier nicht verwendete Operatoren (lineare Weg-Operatoren) aus der Temporallogik:

N a	(Next)	im nächsten Zeitpunkt gilt a
a U b	(Until)	zunächst gilt a, dann b

Possible Worlds in der CTL

b: Cindy ist in Berlin
a: Berlin ist Hauptstadt Deutschlands
c: Cindy ist in Celle
d: es ist Herbst



Multiple Welten resultieren aus dem Mangel an Wissen des Agenten.

Jeder Pfeil entspricht einer Handlung. Im Beispiel hat der Agent anfangs drei Alternativen, die zu den drei Situationen führen. Von dort aus führt jeweils eine Aktion in den nächsten Zustand. Das Wissen des Agenten ist in den Buchstaben an den Situationen kodiert.

a ist zu jedem Zeitpunkt in jeder möglichen Welt gültig, also „unausweichlich immer“. Beispiel: Es ist notwendig (inevitable), daß eins und eins immer (always) zwei ergibt

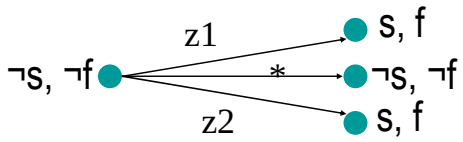
b tritt auf einem Zeitstrahl zu jedem Zeitpunkt auf, ist also „optional immer“. Beispiel: Es ist möglich (optional), daß Mary immer (always) in Australien leben wird

c gilt zu einem bestimmten Zeitpunkt in allen möglichen Welten, ist also „irgendwann unausweichlich“. Beispiel: Es ist notwendig (inevitable), daß die Welt eventuell (eventually) zum Ende kommt

d taucht nur zu einem Zeitpunkt in einer möglichen Welt auf, ist also „optional irgendwann“. Beispiel: Es ist möglich (optional), daß John eventuell (eventually) London besucht.

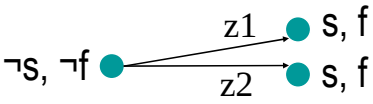
Beispiel für den Einsatz der Possible Worlds Semantik

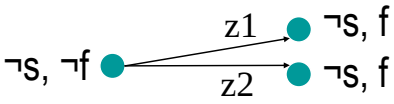
⇨ **Ein Agent glaubt, dass es unvermeidlich ist, dass eine Zahnfüllung (f) durch Schmerz (s) begleitet wird.**



s: Schmerz;
f: Füllung;
z1: Zahnarzt 1;
z2: Zahnarzt 2;
*: andere Aktion

⇨ **Der Agent hat das Ziel, eine Zahnfüllung zu bekommen, aber nicht notwendigerweise das Ziel, Schmerz zu erleiden.**

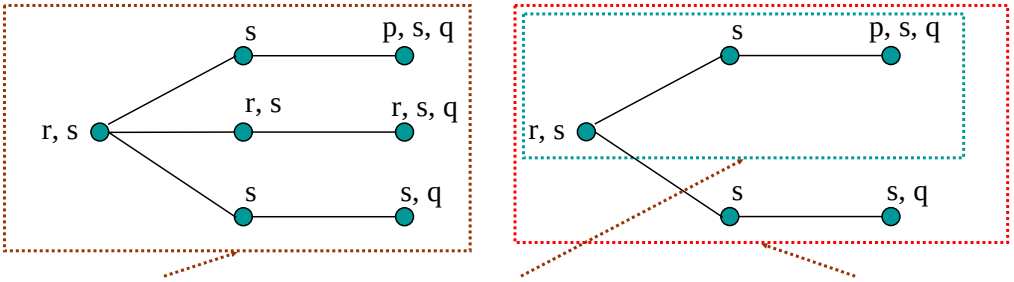




Die Ziele-Welt unten rechts ist nicht konsistent zur Belief-Welt (wegen der Verneinung von Schmerz).

Beziehungen zwischen B, D und I

- ⇨ Goal-Accessible Worlds sind Subwelten der Belief-Accessible Worlds eines Agenten.
- ⇨ Intention Accessible Worlds sind Subwelten der Goal-Accessible Worlds.



Belief-Accessible Worlds Intention-Accessible Worlds Goal-Accessible Worlds

Damit wird ausgedrückt, dass der Agent nur solche Ziele verfolgt, die er auch für erfüllbar hält – in der PWS heisst dies, dass der Zustand (beschrieben mit beliefs) einer der möglichen Weltzustände ist (und damit in der Belief-Welt des Agenten existieren muss). Umgekehrt muss der Agent nicht alle möglichen Belief-Welten erstreben – dies resultiert in der Tatsache, dass die Ziele-Welten typischerweise eine echte Teilmenge der Belief-Welten darstellen.

Intentionen werden auf ähnliche Weise durch Mengen von intention accessible worlds beschrieben

Gliederung

- ⇨ **Einführung in Modallogik**
- ⇨ **Modallogiken für Zeit und Handeln**
- ⇨ **Intentionalität**
- ⇨ **Epistemische Logik mit Belief, Desire, Intention**
  - → Modaloperatoren Belief, Desire, Intention
  - → Axiomatisierungen für B, D, I
- ⇨ **BDI-Agenten**
- ⇨ **Zusammenfassung**

KD45: Axiomatisierung für epistemische Logiken

- ⇨ $Bel\ (p \rightarrow q) \rightarrow (Bel\ p \rightarrow Bel\ q)$ (K)
- ⇨ $Bel\ p \rightarrow \neg Bel\ \neg p$ (D)
- ⇨ $Bel\ p \rightarrow Bel\ Bel\ p$ (4)
- ⇨ $\neg Bel\ p \rightarrow Bel\ \neg Bel\ p$ (5)

Inferenzregeln:

- ⇨ $p \wedge p \rightarrow q \rightarrow q$ (Modus Ponens)
- ⇨ $p \rightarrow Bel\ p$ (Notwendigkeit)

K: If you believe that p implies q then if you believe p then you believe q

D: This is the consistency axiom, stating that if you believe p then you do not believe that p is false

4: If you believe p then you believe that you believe p

5: If you do not believe p then you believe that you do not believe that p is true

Notwendigkeit: you believe all theorems implied by the logic

Schlussfolgerungen in BDI sind nicht trivial

Bel (clever(John) AND John=partner(Sally))

impliziert nicht:

Bel (clever(Sally))

Des (Zahnarztbesuch) AND Bel (Zahnarztbesuch => Schmerz)

impliziert nicht:

Des (Schmerz)

2 Beispiele, die belegen, dass nicht einfach klassische Inferenzregeln gelten.

Axiomatisierungen von Agenten 1

- ⇨ **Wenn der Agent eine Formel beabsichtigt, dann muss er sie als Ziel haben und auch an sie glauben (starker Realismus):**
  - → $Int_a x \Rightarrow Des_a x$ und $Des_a x \Rightarrow Bel_a x$
    - → x steht z.B. für optionally(eventually(Diplom))
  - → (d.h., Agenten haben keine beliebigen Absichten)
- ⇨ **Realismus: Agenten glauben an ihre Intentionen und Ziele**
  - → $Int_a x \Rightarrow Bel_a Int_a x$ und $Des_a x \Rightarrow Bel_a Des_a x$
- ⇨ **schwacher Realismus: Nicht an die Nichterfüllbarkeit glauben**
  - → $Int_a x \Rightarrow \neg Bel_a Int_a \neg x$

Wörtlich übersetzt, ergeben die ersten beiden Formeln wenig Sinn. Stellt man sich aber die Zeitbäume vor, dann ergeben diese Axiome die Submengen-Beziehugnen von Beliefs, Zielen und Intentionen. Mit anderen Worten: Agenten sollen keine beliebigen Absichten verfolgen (so genannter overenthusiastic realism [Cohen & Levesque])

Die Axiome unter dem 2. Strichpunkt definieren eine Art Meta-Wissen: Die Modaloperatoren können nämlich beliebig verschachtelt werden, sodass sich Sätze wie „ich weiss, dass ich weiss, dass ich die Absicht habe....“ formulieren lassen.

Die Begriffe strong realism und weak realism stammen von [Rao & Georgeff]

Weitere Axiome:

Wenn der Agent ein Ereignis beabsichtigt, dann soll er es auch auszuführen versuchen. Dabei steht allerdings nicht fest, ob er dabei auch erfolgreich sein wird. Er wird lediglich versuchen, das Ereignis auszuführen. Dieses Axiom hält den Agenten aber weder davon ab, auch unbeabsichtigte Aktionen durchzuführen noch gibt es Auskunft über verschachtelte Aktionen. Problematisch wird es, wenn mehrer Aktionen zur Auswahl stehen, sie beabsichtigt werden. Man kann dann den Entscheidungsfindungsprozess selbst wiederum als Aktion ansehen oder den Agenten willkürlich entscheiden lassen.

Axiomatisierungen von Agenten 2

- ⇨ **Absichten werden durch Ziele gestützt**
  - → $Int_a x \Rightarrow Des_a Int_a x$
- ⇨ **Bewusstsein bezüglich der Ereignisse**
  - → unabhängig vom Ausgang (Gelingen oder Scheitern)
  - → **done(e) $\Rightarrow$ Bel_a done(e)**
    - → done(e): Ereignis e hat stattgefunden (erfolgreich oder nicht)
- ⇨ **Fallen lassen von Absichten nach endlicher Zeit**
  - → $Int_a x \Rightarrow \mathrm{I}\ E\ \neg Int_a x$

done(e): Ereignisse sorgen dafür, dass von einer Welt in eine andere übergegangen wird. Es ist ein Unterschied, ob ein Ereignis e nicht auftritt oder ob es fehlgeschlagen ist. Beispiel: Klausur nicht bestehen oder nicht zur Klausur gehen.

Je nachdem, welcher Ausgang eintritt, findet sich der Agent nach dem Ausführen einer Aktion in möglicherweise voneinander unterscheidenden Welten wieder.

Verpflichtung (commitment) gegenüber Intentionen

- **Intentionen lösen Handlungen aus:**
  - Aktivierung eines mit der Intention verknüpften Handlungsplans.
- **Entsprechend ist ein Agent gegenüber seinen Intentionen zu Handlungen verpflichtet (committed).**
- **Agenten können Intentionen entweder aufrechterhalten oder verwerfen.**
  - z.B. in Abhängigkeit davon, ob die Handlung erfolgreich durchgeführt wurde.

Agenten sind immer sowohl immer zu ihren Zielen als auch zu den Handlungen verpflichtet.

Drei Arten des Commitments

⇨ **Ein blindly commited Agent verfolgt seine Intentionen, bis er glaubt, sie erreicht zu haben.**

$$Int_a I\, E\varphi \Rightarrow I\, Int_a\, I\, E\varphi\ U\ Bel_a\varphi$$

- Es ist unabdingbar, dass die Intention aufrechterhalten wird, bis („until“) der Agent φ erreicht zu haben glaubt.
- Was geschieht aber, wenn der Agent BEL(¬φ) in seiner Wissensbasis hat?

Erklärung von blindly committed: Der Agent hat die Intention, φ auf jeden Fall irgendwann zu erreichen. Daraus folgt: Es ist unabdingbar, dass diese Intention aufrechterhalten wird bis („until“) der Agent φ erreicht zu haben glaubt.

I = Inevitable (unabdingbar)

U = Until

Drei Arten des Commitments

⇨ **Ein single-minded Agent behält seine Absichten, bis er glaubt, dass er sie erreicht hat oder niemals erreichen wird.**

$$Int_a I\, E\varphi \Rightarrow I\, Int_a I\, E\varphi\ U\ (\ Bel_a \varphi \vee \neg Bel_a O\, E\, \varphi\ )$$

- Die Intention wird fallen gelassen, wenn der Agent nicht mehr an die Erfüllbarkeit des Zustands glaubt.
- Was bedeutet das für ein Ziel $Des_a\ E\ \varphi$?

Single minded definiert die zusätzliche Möglichkeit, die Intention fallen zu lassen, nämlich dann, wenn der Agent nicht mehr an die Erfüllbarkeit des Zustands glaubt.

Still overcommitted to intentions: Never stops to consider whether or not its intentions are appropriate

Modification: stop to determine whether intentions have succeeded or whether they are impossible:
(*Single-minded commitment*)

Drei Arten des Commitments

⇨ **Ein open-minded Agent bleibt seinen Intentionen solange verpflichtet, bis er sie als erreicht ansieht oder er keine entsprechenden Ziele mehr hat.**

$$Int_a I\, E\varphi \Rightarrow I\, Int_a I\, E\varphi\ U\ (\ Bel_a\varphi\ v\ \neg Des_a\ O\ E\ \varphi\ )$$

➙ Die Intention wird auch dann fallen gelassen, wenn das unterstützende Ziel nicht mehr existiert.

Open minded hat als zusätzliche Abbruchbedingung wenn das unterstützende Ziel nicht mehr existiert.

Axiomatisierungen von Agenten 3

- ⇨ **Ist eine Formel notwendigerweise gültig, dann soll der Agent nicht gezwungen sein, sie als Ziel oder Absicht aufzunehmen.**
  - → $Bel_a \mid \varphi \not\Rightarrow Des_a \mid A\ \varphi$

- ⇨ **Ist eine Formel $(\varphi \rightarrow \psi)$ notwendigerweise gültig und hat der Agent das Ziel (oder die Absicht) φ, dann soll er nicht gezwungen sein, auch b als Ziel (oder Absicht) zu haben.**
  - → $Bel_a \mid (\varphi \rightarrow \psi) \wedge Des_a \mid \varphi \not\Rightarrow Des_a \mid \psi$

Notwendigerweise meint immer oder irgendwann.

1.: z.B. 1+1=2 macht keinen Sinn als Ziel aufzustellen, genauso wenig wie die Absicht zu haben, dass die Sonne aufgeht (tut sie irgendwann auf jeden Fall).

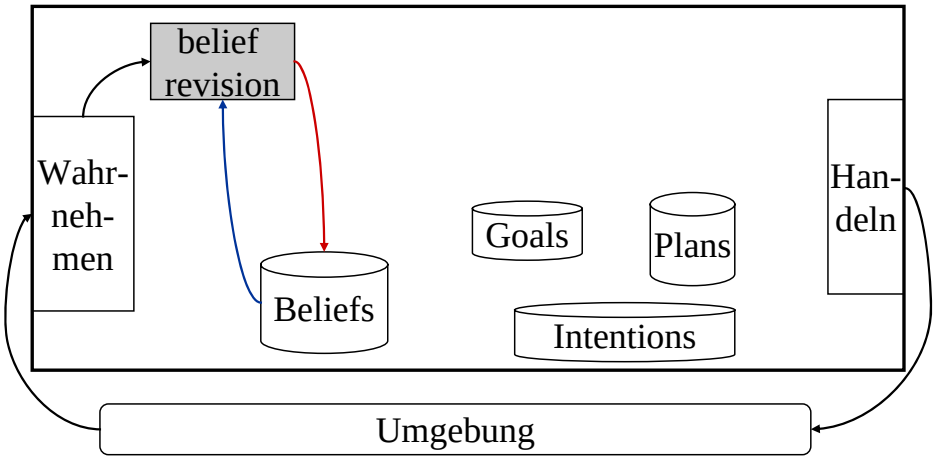
2. Eine andere Interpretation dieses Axioms ist, dass, ein Agent nicht immer alle Seiteneffekte einer Aktion zu berücksichtigen hat. Beispiel: Zahnarzt (der Besuch ist mit Schmerzen verbunden, die aber nicht Ziel des Agenten sind (der Besuch dagegen schon).

Gliederung

- ⇨ **Einführung in Modallogik**
- ⇨ **Modallogiken für Zeit und Handeln**
- ⇨ **Intentionalität**
- ⇨ **Epistemische Logik mit Belief, Desire, Intention**
- ⇨ **BDI-Agenten**
  - → Allgemeine BDI-Architektur
  - → Agent-0
- ⇨ **Zusammenfassung**

Modell einer BDI-Agentenarchitektur

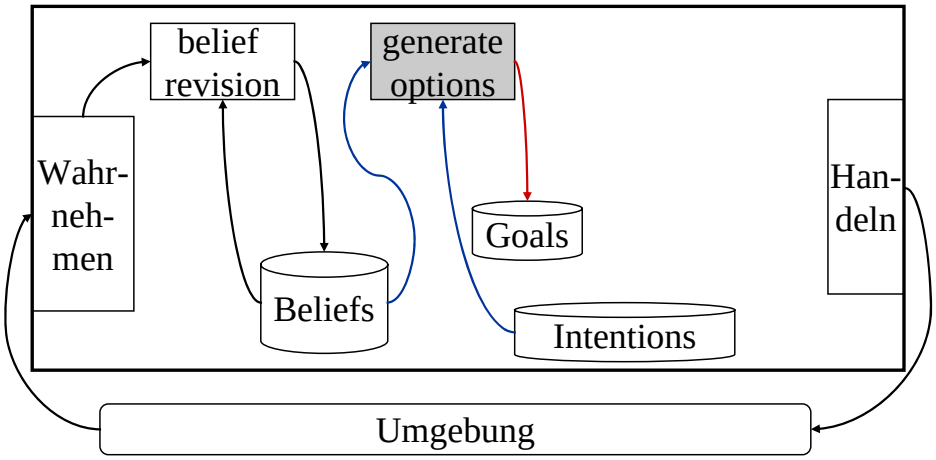
1. Belief revision function: Bel* x Perception → Bel*



0. Die Standard-BDI-Architektur besteht aus Wissensbasen für Beliefs, Goals und Intentions und einer Planbibliothek, die die Handlungspläne enthält. Die Perzeptoren und Effektoren kennen wir von den bereits bekannten Architekturschemata.
1. Wahrnehmungen werden durch die Wissensaktualisierung (belief revision), auf Basis des aktuellen Wissens aktualisiert. Diese Funktion führt zu einem Update der Belief-Wissensbasis.

Modell einer BDI-Agentenarchitektur

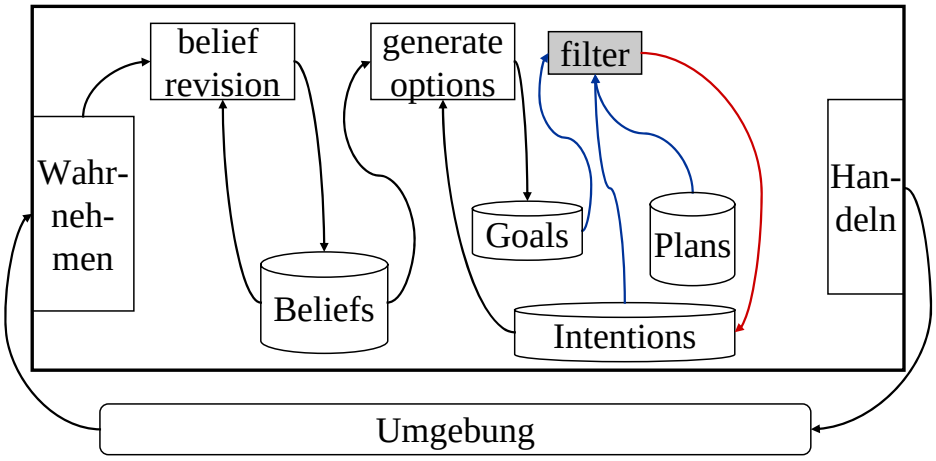
2. Option generation : Bel* x Int* → Goal*



2. Aus Beliefs und unter Beachtung der aktuellen Intentionen werden neue Ziele generiert und nicht mehr aktuelle Ziele entfernt. In der Praxis ist auch die Zieldatenbasis Eingabeparameter der „generate options“-Funktion, d.h.: generate options: Bel* x Goal* x Int* → Goal*.

Modell einer BDI-Agentenarchitektur

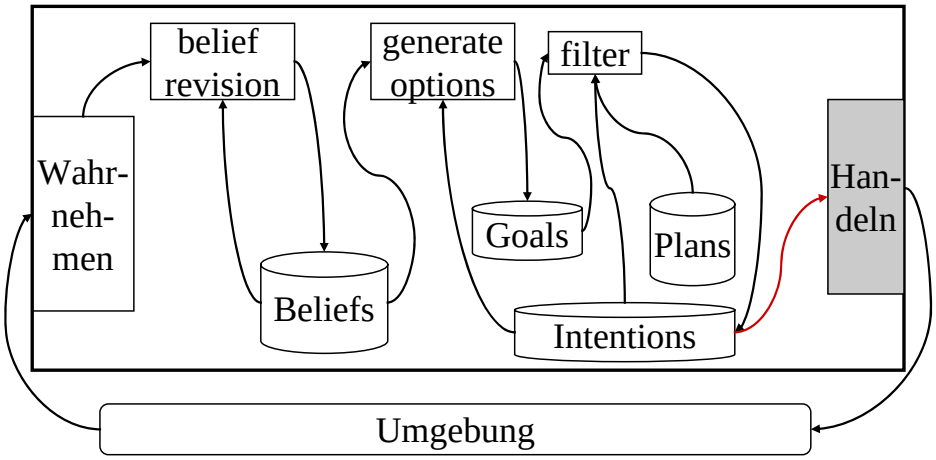
3. filter : Goal* x Int* x Plan* → Int*



3. Aus den Zielen werden unter Beachtung der Intentionen neue Pläne instantiiert und in die Intentionenstruktur eingetragen. Nicht länger gültige (z.B. weil erfüllte) Ziele führen dazu, dass Intentionen fallen gelassen werden.

Modell einer BDI-Agentenarchitektur

4. ausführen: Int* → Action



4. Intentionen werden ausgeführt. Wenn ein Plan aus einer Sequenz von Handlungen besteht (typische Annahme), werden die einzelnen Handlungsschritte nacheinander der Handlungskomponente übergeben.

Gliederung

- Einführung in Modallogik
- Modallogiken für Zeit und Handeln
- Intentionalität
- Epistemische Logik mit Belief, Desire, Intention
- BDI-Agenten
  - Allgemeine BDI-Architektur
  - Agent-0
- Zusammenfassung

AGENT-0 [Shoham 1993]

Agent-Oriented Programming besteht nach Shoham aus:

- ⇨ **logischem System zur Beschreibung mentaler Zustände,**
- ⇨ **interpretierter Programmiersprache für Agenten sowie**
- ⇨ **Agentifizierungsprozess: Kette von Übersetzungsschritten.**

AGENT-0: Logisches System

AGENT-0 unterstützt folgende Sprachelemente

- ⇨ **belief (mental) - BEL**
- ⇨ **commitment (mental) - CMT**
- ⇨ **capability (nicht mental) - CAN**

- ⇨ **Grammatik = modale Prädikatenlogik, erweitert um Zeitpunkte**

mentaler Belief-Operator

BEL(<agent>, <timepoint>, <fact>)

BEL(a, t1, BEL(b, t2, like(a, b, t3)))

„Agent *a* glaubt zum Zeitpunkt *t1*, dass Agent *b* zum Zeitpunkt *t2* an *like(a,b,t3)* glaubt“

nicht-mentaler Capability-Operator

CAN(<agent>, <timepoint>, <fact>)

CAN(a, t1, open(door, t2))

Fähigkeit wird durch Zustand ausgedrückt

„Agent *a* kann zum Zeitpunkt *t1* sicherstellen, dass die Tür zum Zeitpunkt *t2* offen ist“

Mentaler Commitment-Operator

CMT(<agent>, <agent>, <timepoint>, <fact>)

CMT(a, b, t1, open(door, t2))

„Agent *a* ist Agent *b* zum Zeitpunkt *t1* zur Türöffnung verpflichtet“

Logisches System: Annahmen

- Konsistenz von Beliefs und Commitments

- Good faith:
  CMT(a,b,t,x) <=> BEL(a, t, CAN(a, t, x))

- Introspection:
  CMT(a,b,t,x) <=> BEL(a, t, CMT(a, b, t, x))

  ¬CMT(a,b,t,x) <=> BEL(a ,t, ¬CMT(a, b, t, x))

Die AGENT-0 -Sprache:
communicative actions und private actions

⇨ **inform(t, a, fact)**

➙ zum Zeitpunkt *t* dem Agenten *a* das *fact* zusenden

⇨ **request(t, a, action)**

➙ Agent *a* zu Handlung *action* auffordern

➙ unrequest und refrain analog.

⇨ **DO(t, p-action)**

➙ zum Zeitpunkt *t* die lokale Methode *p-action* ausführen

inform: zum Zeitpunkt t dem Agenten a das Fakt fact zusenden.
DO: zum Zeitpunkt t die private Aktion p-action ausführen.
zusätzliche communicative actions:

refrain(t, a, action)	**Anfrage zum Nichthandeln**
unrequest(t, a, action)	**Rücknahme eines request**

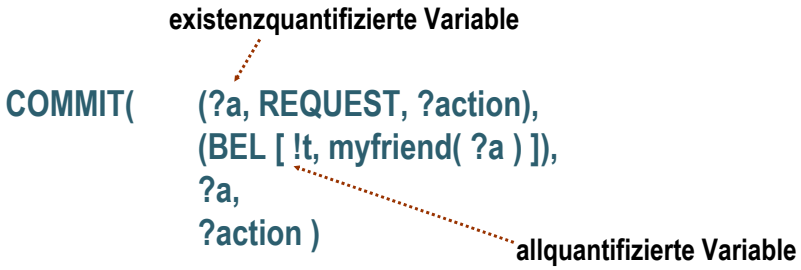
(mehr Folien im Anhang)

Die AGENT-0 -Sprache:
Commitment Rules

COMMIT(msgcond, mtlcond, agent, action)

msgcond = (From, Type, Content)

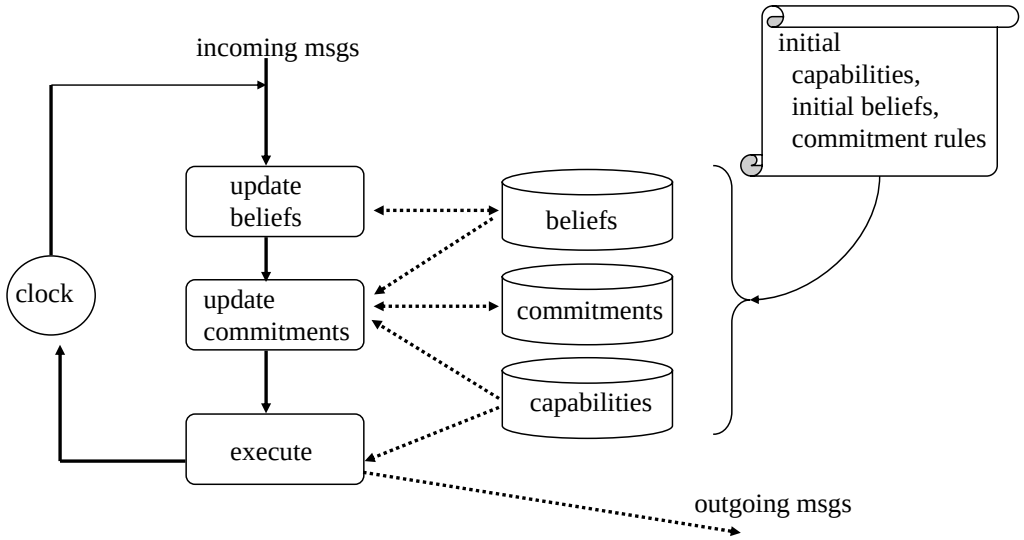
„aufgrund einer empfangenen Nachricht in einem bestimmten Zustand einem anderen Agenten zu einer Aktion verpflichtet sein."



„requests von Freunden werden ausgeführt"

!t ist allquantifizierte Variable, bedeutet dass myfriend(?a) für alle Zeit gilt.

Agent-0: Architektur und Interpreterschleife



Ein Agent besitzt 3 Wissensbasen, die durch das initiale Agentenprogramm gefüllt werden. Ziele fehlen, Intentionen sind Commitments, Pläne sind als capabilities notiert.

Durch Uhr gesteuerte Interpreter-Schleife: Empfangene Nachrichten führen dazu, dass die Belief-Wissensbasis aktualisiert wird. Aus neuen Beliefs werden die Commitments aktualisiert. Commitments schließlich triggern Aktionen, die ausgeführt werden.

Agent-0: Interpreterschleife: 1. Belief-Update

betrifft inform-Sprechakte:

inform(t1, a, fact(x, t2))

- Einfügen des Fakts in die Belief-Datenbasis
- Entfernen inkonsistenter Fakten

Welche Fakten werden entfernt? Solche, die Negationen zu den empfangenen Informationen darstellen und alle diejenigen Fakten, deren Gültigkeit abgelaufen ist.

Agent-0: Interpreterschleife: 2. Commitment-Update

commit(msgcond,mtlcond,a,action)

Die Verpflichtung wird angenommen, wenn

a) msgcond mappt msg,

b) mtlcond ist wahr,

c) action ist aktuelle capability,

d) Agent ist nicht zu REFRAIN(action) verpflichtet.

unrequest(action): Entfernen der Handlung aus Commitment-DB

Ein unrequest(action) führt zum Entfernen der Handlung aus Commitment-DB.
msgcond ist eine message condition (Mapping eines empfangenen Sprechakts)
mtlcond ist eine mental condition (Mapping auf aktuelle Beliefs)

Agent-0: Interpreterschleife: 3. Execute

3. Ausführung von Commitments in Abhängigkeit der Zeit:

INFORM(t, b, fact)

a) senden

b) assert(BEL(t,a, BEL(t,b,fact)))

REQUEST, UNREQUEST

senden

DO(t, action)

a) action checken,

b) ggfs. ausführen

Gliederung

- ⇨ Einführung in Modallogik
- ⇨ Modallogiken für Zeit und Handeln
- ⇨ Epistemische Logiken
- ⇨ Zusammenfassung

Modallogiken — Zusammenfassung

	Syntax	Inferenz	Semantik
Modal	Möglichkeit ◊ und Notwendigkeit □	Inferenzregeln für ◊ und □	Possible Worlds
Dynamisch	Sequenz, Verzweigung, Test	Inferenzregeln für Ergebnisse	Possible Worlds mit Übergängen
Temporal	Zeitpunkte oder Intervalle, zeitliche Beziehungen	Inferenz für Aussagen mit zeitlicher Ergänzung	Possible Worlds mit vielfältigen Übergängen

Modalitäten sind Qualifikationen von Propositionen. Ein Modaloperator steht vor der Proposition und ändert deren Bedeutung.

Zeitlogik: typischerweise immer (always), irgendwann (eventually), folgt direkt (meets) und until.

Zusammenfassung Intentionalität

- ⇨ **Der intentional stance bietet eine pragmatische und natürliche Methode, die Regeln und Verhaltensmuster unabhängig von der tatsächlichen Implementation erkennen lässt.**
- ⇨ **Es geht nicht darum, ob ein Agent wirklich etwas „glaubt“, „wünscht“ oder „beabsichtigt“.**
- ⇨ **BDI erweist sich als besonders nützlich, wenn ein Agent über andere Agenten „nachdenken“ soll.**

Intentionalität bezieht sich auf einzelne Agenten, nicht auf Agentengemeinschaften wie es z.B. die Koordination tut.

BDI-Logiken: Zusammenfassung

- ⇨ **Der Agent wird als intentionales System beschrieben**
- ⇨ **Modallogik wird für nicht-monotones Schließen verwendet**
  - → Modaloperatoren Bel, Des, Int
  - → verschiedene Axiomatisierungen sind möglich
- ⇨ **BDI überbrückt die Lücke zwischen Wissen und Handeln**
  - → Aktionen werden über Intentionen,
    Intentionen über Ziele und
    Ziele über Wissen motiviert.

Monotones Schließen: Kenntnis weiterer Fakten erhält alle bisherigen Ableitungen, d.h. neues Wissen kann nicht kontradiktorisch zu bestehendem sein und kann auch altes Wissen nicht überschreiben.

Der Weg der Information durch den Agenten: Aus der Umwelt kommen Wahrnehmungen, die in der Belief-Wissensbasis verwaltet werden und aus denen Ziele generiert werden. Ziele führen zur Handlungsauswahl (also Intentionen) und unmittelbar nachfolgend zur Modifikation der Umwelt durch Aktionen.

Kritik an BDI-Architekturen:

Sprechaktkommunikation und insbesondere Interaktionsprotokolle spielen kaum eine Rolle (Ausnahmen wie AgentBuilder und JACK (nächste VL) bestätigen die Regel). Somit wird die Koordination von Multiagentensytemen mit BDI-Architekturen nicht unterstützt.

Schlechte Lernfähigkeit: In der Architektur sind Wahrnehmen, Räsonnieren und Handeln relativ fest verdrahtet - wo soll Lernen ansetzen? Außerdem bietet BDI keinen Mechanismus, um die Effekte der eigenen Aktionen zu erfahren und somit durch Feedback der Umgebung zu lernen.

Mehr zu Lernen in der nächsten Vorlesung über „intelligente Agenten“

Vorteile einer *expliziten Repräsentation* von Zielen in Agenten

- **höhere Fehlertoleranz:**
  - Schlägt eine Handlung fehl, kann der Agent besser wieder aufsetzen.
- **bessere Koordination:**
  - Widersprüchliche Ziele nicht gleichzeitig verfolgen.
  - Subsumierende Ziele nur einmal verfolgen.
  - Ständiges Überdenken und Umplanen bei veränderten Randbedingungen ist möglich.

BDI-Logiken: Diskussion und Ausblick

- ⇨ **Welche Modaloperatoren verwenden?**
  - → Wissen, Know How, Kommunikation
  - → Commitment, Joint Goal, Joint Commitment
- ⇨ **Die Axiomatisierungen sind problematisch**
  - → Constraintformalismen drücken keine Aktivitäten aus
  - → Wo kommen Ziele her?
- ⇨ **Der Umgang mit Widersprüchen und veralteten Informationen ist unklar.**

Eine Reihe von Fragen werfen BDI-Logiken auf:

Sind die verwendeten Anthropomorphismen statthaft? Ist es grundsätzlich erlaubt und sinnvoll, einem Software-Programm Wissen, Ziele und Intentionen zuzuschreiben?

Singh has developed a logic for representing intentions, beliefs, knowledge, know-how, and communication in a branching-time framework. Cohen und Levesque definieren formallogisch die Begriffe Commitment, Joint Goal, Joint Commitment.

Constraintformalismen drücken keine Aktivitäten aus, sie beschreiben nur gültige Zustände, nicht aber, wie man einen Agenten in diese Zustände überführt. Beispiel Beliefs: wo kommen sie her, wie werden sie aktualisiert?

Weitere Probleme klassischer Logik — Zusammenfassung

Klassische Logik ist als Repräsentationsformalismus ungeeignet für viele Aspekte der wirklichen Welt:

- ⇨ **zur Repräsentation kontinuierlicher Größen,**
- ⇨ **zur Repräsentation des Unbekannten,**
- ⇨ **zur Repräsentation unsicheren, probabilistischen Wissens,**
  - → „morgen ist 80%ige Regenwahrscheinlichkeit“
- ⇨ **zur Repräsentation von Relativität und Unschärfe.**
  - → „Paul ist ziemlich groß“.

Logik sowohl als Repräsentationsformalismus als auch als Inferenzmechanismus.

kontinuierlicher Größen/Werte: z.B. Maße, Farben.

„Paul ist ziemlich groß“. Wenn auch Petra ziemlich groß ist, sind dann Paul und Petra gleich groß?

Monotonie: Wenn eine Aussage H aus X abgeleitet werden kann, dann auch auch ♋◆♦ ✠ ◡ Y

D.h., neues Wissen f□hrt niemals dazu, dass altes Wissen nicht mehr gilt.


DAI-Labor
TU Berlin
Grundlagen der
Künstlichen Intelligenz
08.12.2005: Modallogik
Abschluss der ersten Teils GKI
„symbolische KI“
ab nächste Woche: statistische
Verfahren
Dr.-Ing. Stefan Fricke
stefan.fricke@dai-labor.de
AOT
Agententechnologien in
betrieblichen Anwendungen
und der Telekommunikation

Lernziele:

Referenzen

⇨**Agent-0:**

- ➙Y. Shoham. Agent-oriented programming. Artificial Intelligence, 60. S. 51 - 92, 1993.
- ➙www.agentbuilder.com: kommerzielles Agent-0

⇨**BDI:**

- ➙Modeling Rational Agents within a BDI – Architecture (Anand. S. Rao, Michael P. Georgeff)
- ➙Andreas Kerlin, Anatolij Zubow, Daniel Göhring: Glauben und Absichten. Seminararbeit HU Berlin, 2002, http://www.drgoehring.de/uni/papers/Glauben_und_Absichten_062002.pdf

Anhänge

BDI-Programm

```
Beliefs = Beliefs_0
Intentions = Intentions_0  Goals = Goals_0
while  true  do
        get next perceipt p
        Beliefs = brf(Beliefs,p)
        Intentions = options(Goals,Intentions)
        Intentions = filter(Beliefs, Goals, Intentions)
        π = plan(Beliefs, Intentions)
        while not (empty(π) or succeeded (Intentions, Beliefs)  or
                    impossible(Intentions, Beliefs))  do
                α = head(π)
                execute(α)
                π = tail(π)
                get next perceipt p
                Beliefs = brf(Beliefs,p)
                if not sound(π, Intentions, Beliefs) then
                    π = plan(Beliefs, Intentions)
        end while
end while
```

Dropping intentions that are impossible or have succeeded

Reactivity, replan

Schwierige Anwendbarkeit der Prädikatenlogik für bestimmte Umgebungstypen:

- ⇨ **Nichtdeterministisch** → **unzuverlässige Aktionen**
- ⇨ **Partiell beobachtbar** → **unvollständiges Wissen**
- ⇨ **Kontinuierlich** → **Repräsentationsproblem**
- ⇨ **Dynamisch** → **unzuverlässiges Wissen**

Die Eigenschaft episodisch bzw. sequentiell hat keinen Einfluss.

Die Eigenschaft single/multi agent bringt keine neuen Probleme zusätzlich zu den oben genannten (insbesondere durch Nichtdeterminismus, Dynamik).

Agent-0 – Zusammenfassung

- ⇨ **Der Agent ist im Wesentlichen ein Regelinterpreter.**
- ⇨ **Ziele werden nicht modelliert.**
- ⇨ **Zeitpunkte sind problematisch im Zusammenhang mit Wissen.**
- ⇨ **Wenig Aufmerksamkeit wird dem belief-revision geschenkt.**
- ⇨ **AgentBuilder ist eine kommerzielle Erweiterung von Agent-0**
  - → mit Sprechakten und Interaktionsprotokollen
  - → www.agentbuilder.com

Commitment-Regeln sind das Herz eines Agenten.

Es gibt mächtigere Zeitlogiken, in denen Operatoren für immer, später, etc. existieren.