Ist künstliche Motivation gefährlich?

„Das Problem ist gar nicht künstliche Intelligenz, sondern künstliche Motivation.“

Referent: Marvin Schmitt

Dozent: PD Dr. Ullrich Köthe

Mitsukus Motivation

Me: Are you *intelligent*?

Bot: I am ***highly intelligent***. I have many intellectual functions

Me: For example?

Bot: ***Reasoning, thinking, deduction*** and ***self-awareness***, to name a few.

Me: Do you have any *motivation*?

Bot: ***Where would I get it***? But I do have an answer to every question..

Gliederung

1 **Ist Intelligenz alles?**
   1 Was ist ein Genie?
   2 Prädiktoren für Erfolg

2 **Motivation in Psychologie**
   1 Motiv vs. Motivation
   2 Aktivationstheorie nach Berlyne
   3 Risiko-Wahl-Modell nach Atkinson
   4 Studie im Seminar

3 **Motivation in der KI-Forschung**
   1 Orthogonalitätsthese
   2 Ethik und Verantwortung

4 **Fazit und Ausblick**

1 Ist Intelligenz alles?

Def. Intelligenz ist das, was ein Intelligenztest misst. *(Boring, 1923)*

- Ist die Leistung in einem Intelligenztest das, was uns wirklich interessiert?
- Ist die ideale KI „einfach nur intelligent“ oder gibt es noch andere Eigenschaften, die wichtig sind?
- Begriff des ***„Erfolgs“*** → Umsetzung von Fähigkeiten in tatsächliche Handlungen

1.1 Was ist ein Genie?

nach Simonton (2016)

Def. *„A person who has an exceptionally high Intelligence Quotient (IQ), typically above 140.“*

$$IQ = \frac{Lebensalter}{Intelligenzalter} \cdot 100$$

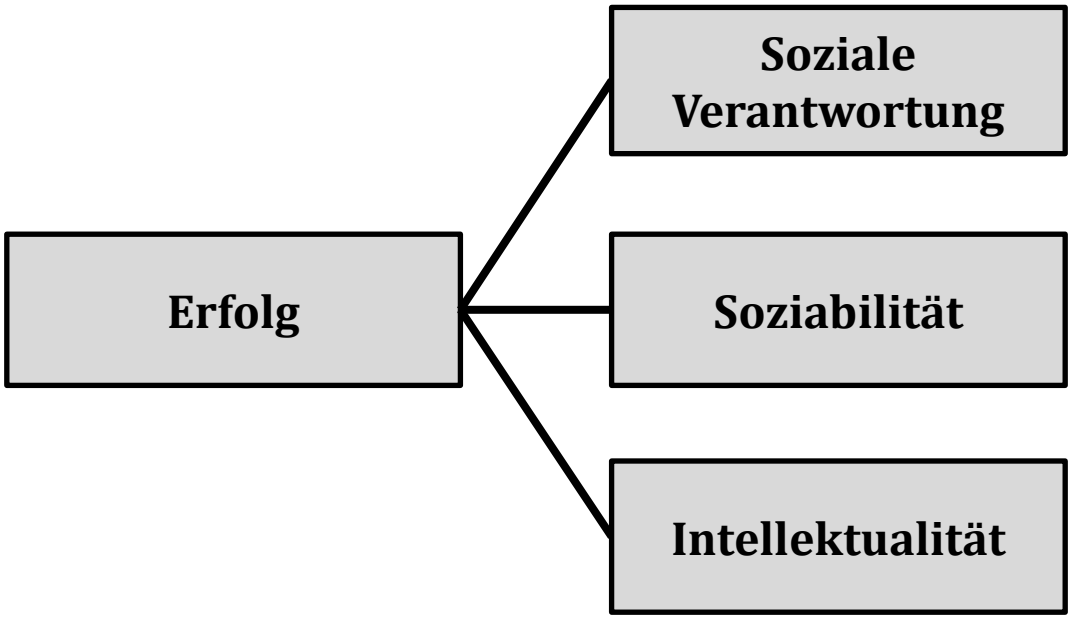
- Herkunft: **Terman-Studie** der intelligentesten 1% Kinder, die er als „Genies“ beschrieb
- Methode der Terman-Studie: Historiometrische Analyse

Def. *Idiographische Methode, um Leistungs- und Charaktermerkmale herausragender Persönlichkeiten unter Rückgriff auf biographisches Material zu quantifizieren und zu vergleichen.*

- Ergebnisse der Terman-Studie: Identifikation verschiedener Prädiktoren für Erfolg

1.2 Prädiktoren für Erfolg

nach der Terman-Studie



z.B. Gewissenhaftigkeit, Aufrichtigkeit

z.B. Popularität, Offenheit, Liebenswürdigkeit, familiäre Harmonie *(overall bester Prädiktor!)*

z.B. allgemeine Intelligenz, Common Sense, Wissensdrang

Intelligenz ist nur ein Prädiktor von vielen!

Analyse von Cox (1926): Intelligenz erklärt nur *10% der Varianz* von Erfolg

→ Suche nach weiteren Prädiktoren für Erfolg ist notwendig!

1.2 Prädiktoren für Erfolg

nach Cox (1926)

„ […] that high but not the highest intelligence, combined with the greatest degree of [motivational] persistence, will archieve greater eminence than the highest degree of intelligence with somewhat less [motivational] persistence.“ (Cox, 1926)

→ Intelligenz als notwendiges aber nicht hinreichendes Kriterium

→ Betont die Rolle von Motivation bzw. Ausdauer

- → In der modernen Forschung auch als „Grit“ bezeichnet
- → Grit klärt über Intelligenz hinaus 4% an Varianz daran auf, ob man langfristige Ziele erreicht (Matthews and Kelly, 2007)
- → Mechanismus ist jedoch noch unklar, wird mit Mediatoranalysen untersucht

2 Motivation in der Psychologie

- Allgemein: Fokus auf zugrundeliegende Motive und die tatsächliche Realisierung in konkreten Situationen

Was ist Motivation in der Psychologie nicht?

- Motivation ist nicht gleich „ich bin voll motiviert, ey“!
- Es gibt keine „gute Motivation“ vs. „schlechte Motivation“

Wichtige Unterscheidungen

- Motiv vs. Motivation
- Personismus vs. Situationismus vs. Interaktionismus

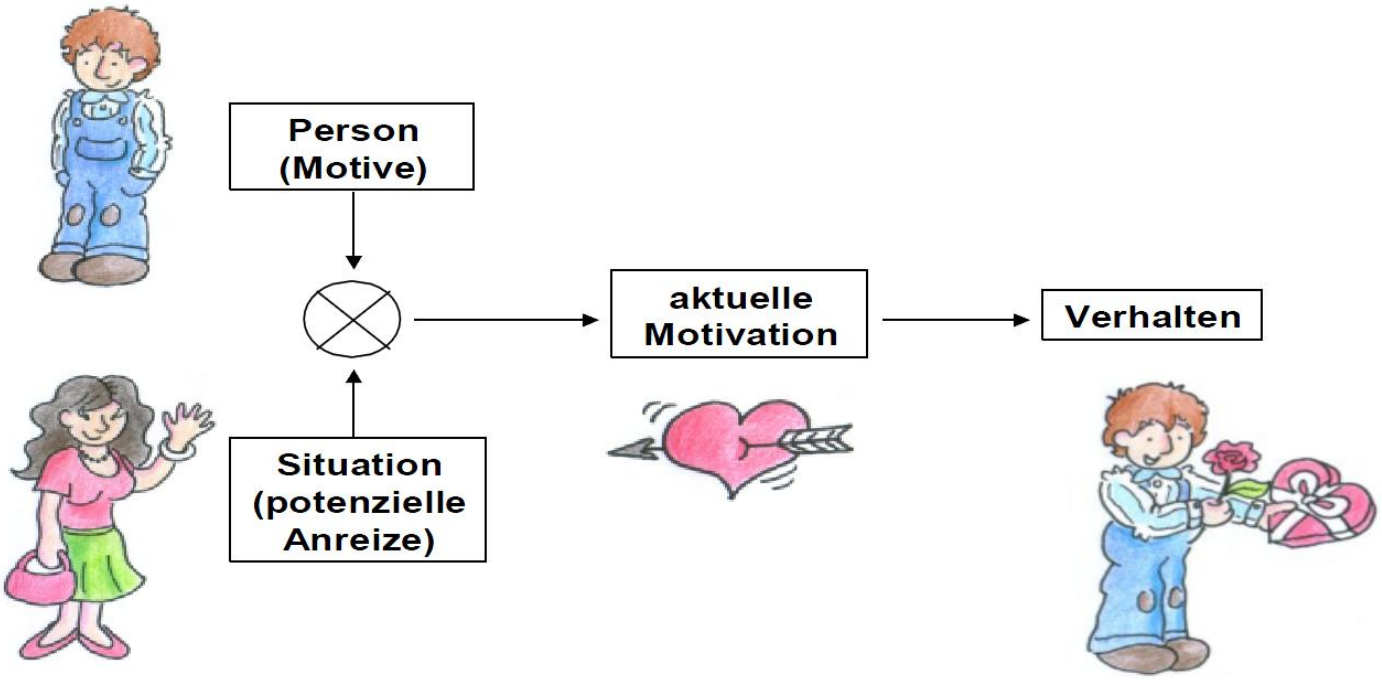
2.1 Motiv vs. Motivation

Motiv	Motivation
„Überdauernde Vorlieben einer Personen, die sich auf inhaltliche Klassen von Handlungszielen beziehen“ (Heckhausen, 1989)	*Aktualisierung eines Motivs in einer konkreten Situation*
Funktion:	
→Verbindung zwischen Biologie und Verhalten → Zuweisung von Verantwortlichkeit → Schluss von offenem Verhalten auf innere Zustände →Erklärung von Verhaltensvariabilität	→Beschreibt den Zustand eines Organismus, ein Ziel aufzusuchen oder zu vermeiden (approach vs. avoidance) →Dimensionen: Wahl, Intensität, Latenz und Persistenz

<u>Frage:</u> Führt ein Motiv zwangsweise zur Motivation?

2.1 Motiv vs. Motivation

Motivation als Produkt aus Motiven und Situation



Übernommen von Prof. Dr. Ursula Christmann, 2014, PI Heidelberg
mod. nach Rheinberg, F. (2004). Motivation. Stuttgart: Kohlhammer, S. 70

2.2 Aktivationstheorie

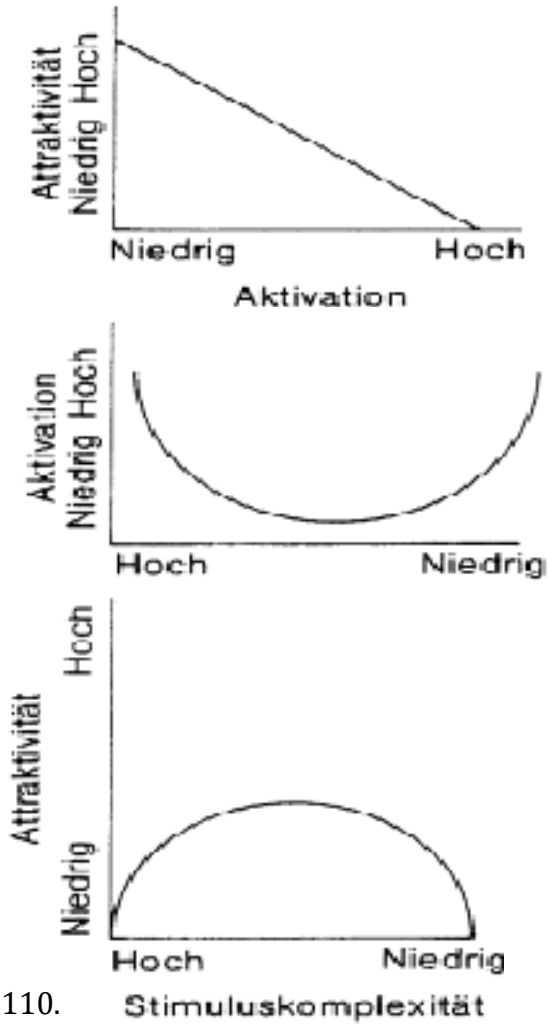
nach Berlyne

1. ***Niedrige Aktivation*** führt zu
   Hoher Attraktivität

&

2. ***Mittlere Komplexität*** führt zu
   niedriger Aktivation

3. ***Mittlere Komplexität*** führt zu
   hoher Akttraktivität

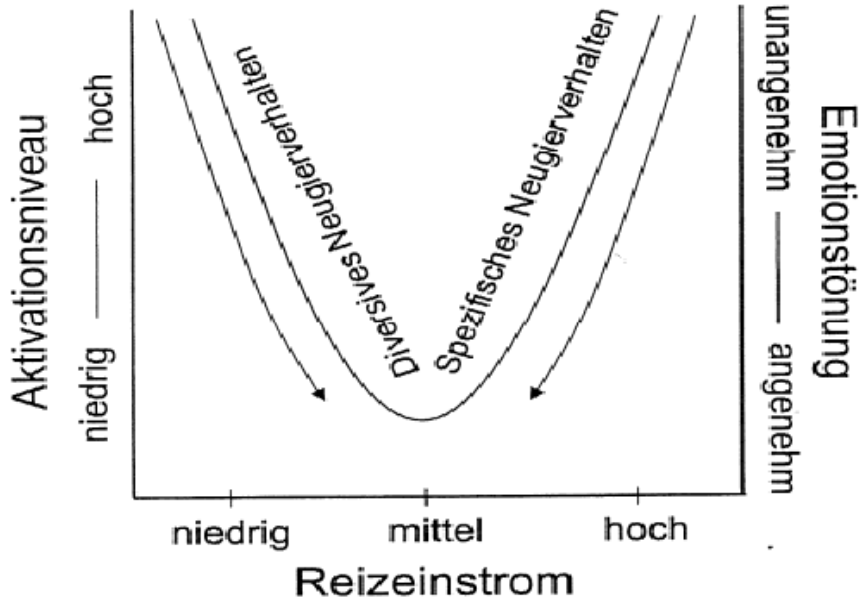


Grafiken aus: Weiner, Bernard (1994). *Motivationspsychologie*. Weinheim: Beltz. 110.

2.2 Aktivationstheorie

nach Berlyne

Erklärung von Neugierverhalten



→ Regression zur Mitte: Reizeinstrom auf mittleres Niveau bringen um die Attraktivität zu maximieren

2.2 Aktivationstheorie

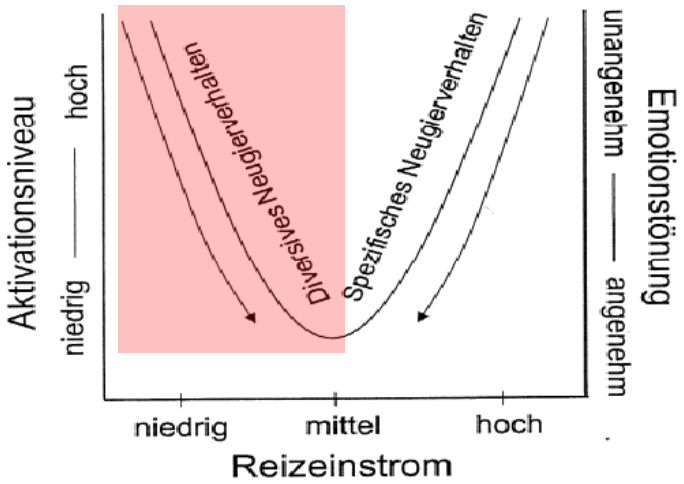
nach Berlyne

<u>Diversives Neugierverhalten</u>

→ Erhöhung des Reizeinstroms

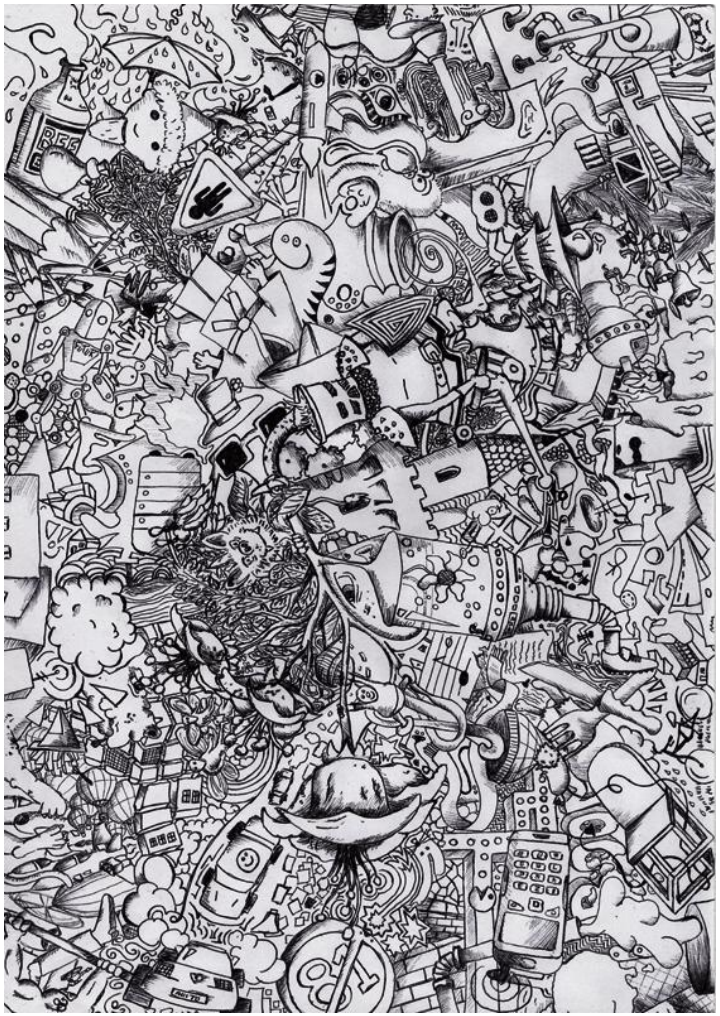
→ Hier: Interpretation

→ Beispiel: Fernsehkanäle „zappen“



2.2 Aktivationstheorie

nach Berlyne

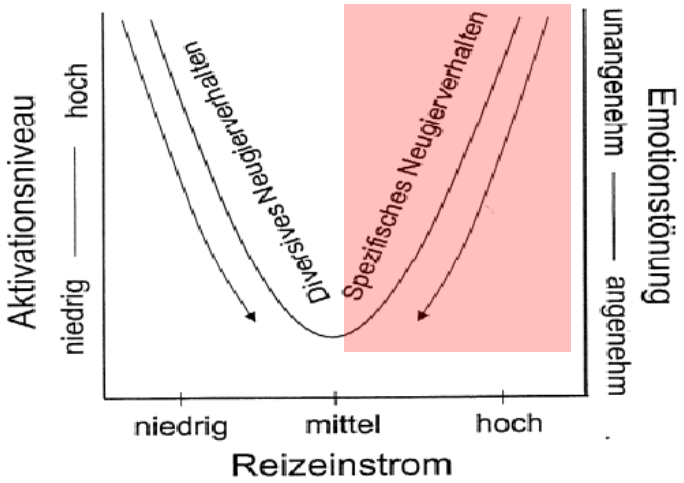


<u>Spezifisches Neugierverhalten</u>

→ Exploration

→ Reduktion des Reizeinstroms durch Fokus auf kleine Bereiche

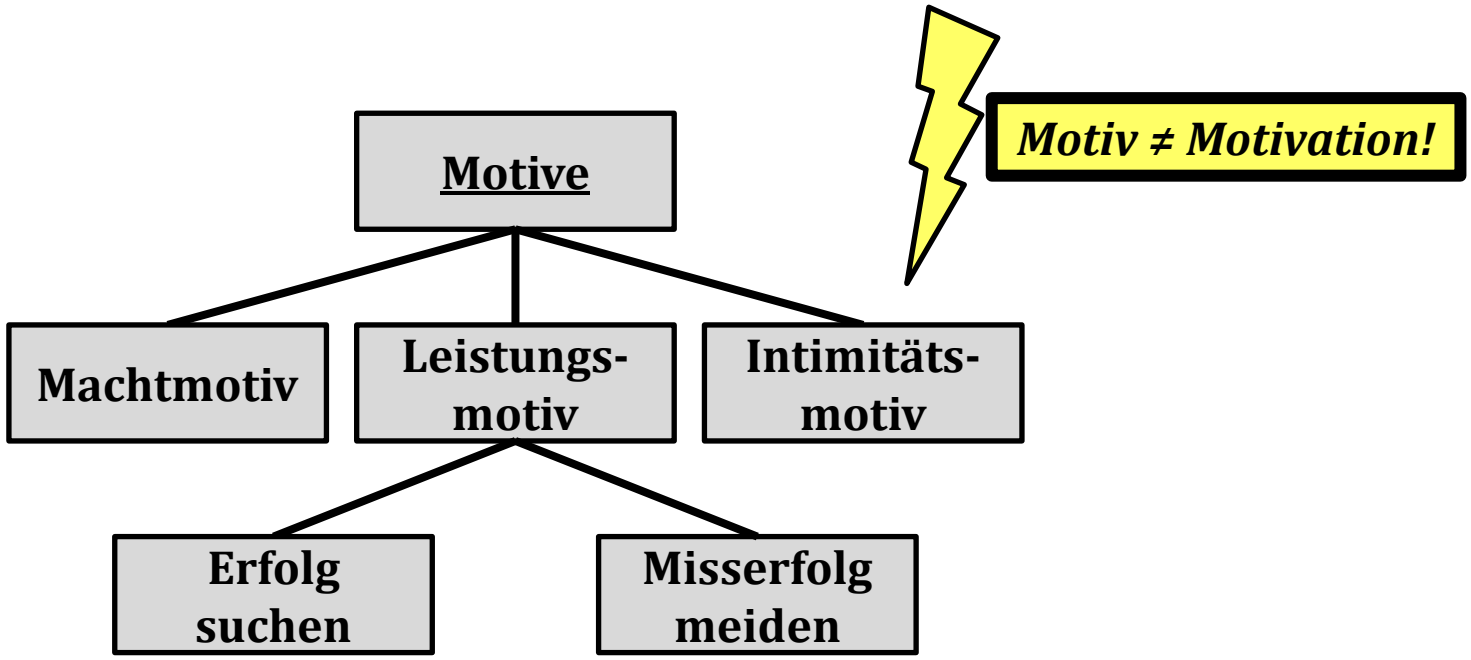
→ Beispiel: Arbeit mit einem Kursbuch



2.3 Risiko-Wahl-Modell

nach Atkinson (1957)

Motive Disposition Theory (McClelland)

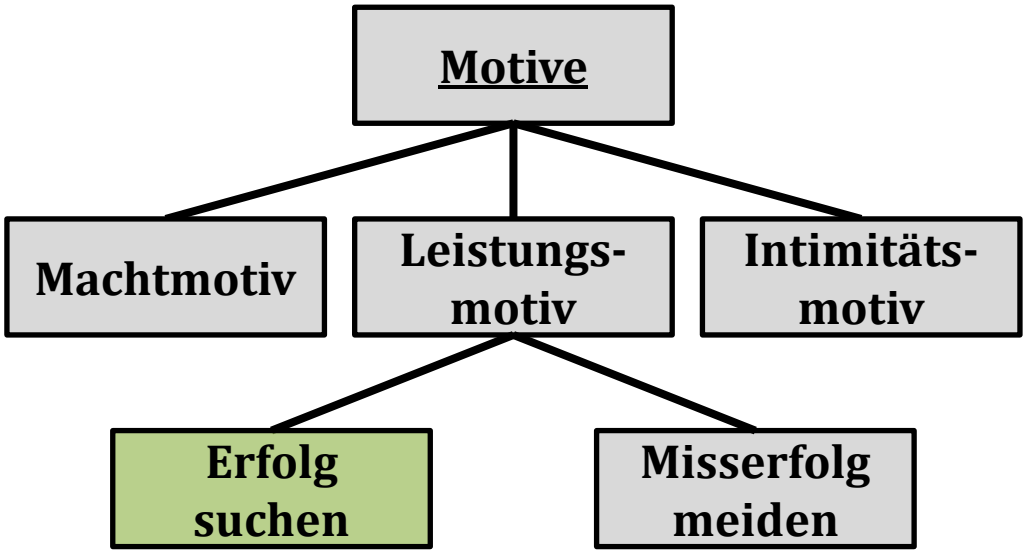


→ Nach McClelland et al. (1953) sind die drei *Hauptmotive* das *Machtmotiv*, das *Leistungsmotiv* und das *Intimitäts-* bzw. *Zugehörigkeitsmotiv*.

→ Nach Atkinson (1957) besteht das *Leistungsmotiv* aus der *Tendenz, Erfolg zu ersuchen* und der *Tendenz, Misserfolg zu vermeiden*.

2.3 Risiko-Wahl-Modell

nach Atkinson (1957)



Tendenz, Erfolg zu suchen: „Fähigkeit, stolz zu sein“

$$T_E = M_E \cdot W_E \cdot A_E$$

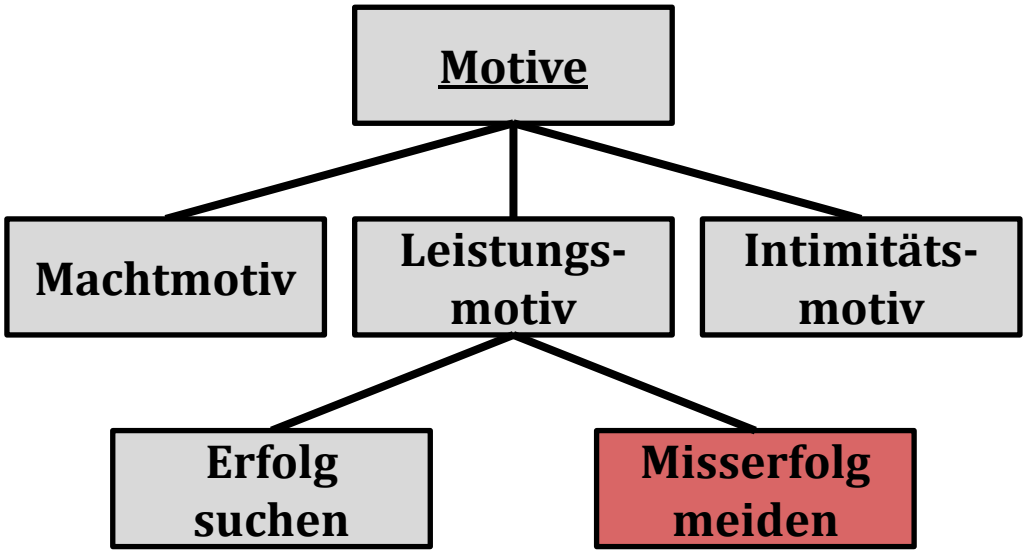
Erfolgsmotiv

Subjektive Erfolgswahrscheinlichkeit

Anreiz von Erfolg = 1-W_E → Je schwieriger die Aufgabe, desto mehr Stolz

2.3 Risiko-Wahl-Modell

nach Atkinson (1957)



Tendenz, Misserfolg zu meiden:
„Mit Scham reagieren“

$$T_M = M_M \cdot W_M \cdot A_M$$

Misserfolgsmotiv

Subjektive Misserfolgswahrscheinlichkeit

Anreiz von Misserfolg

2.3 Risiko-Wahl-Modell

nach Atkinson (1957)

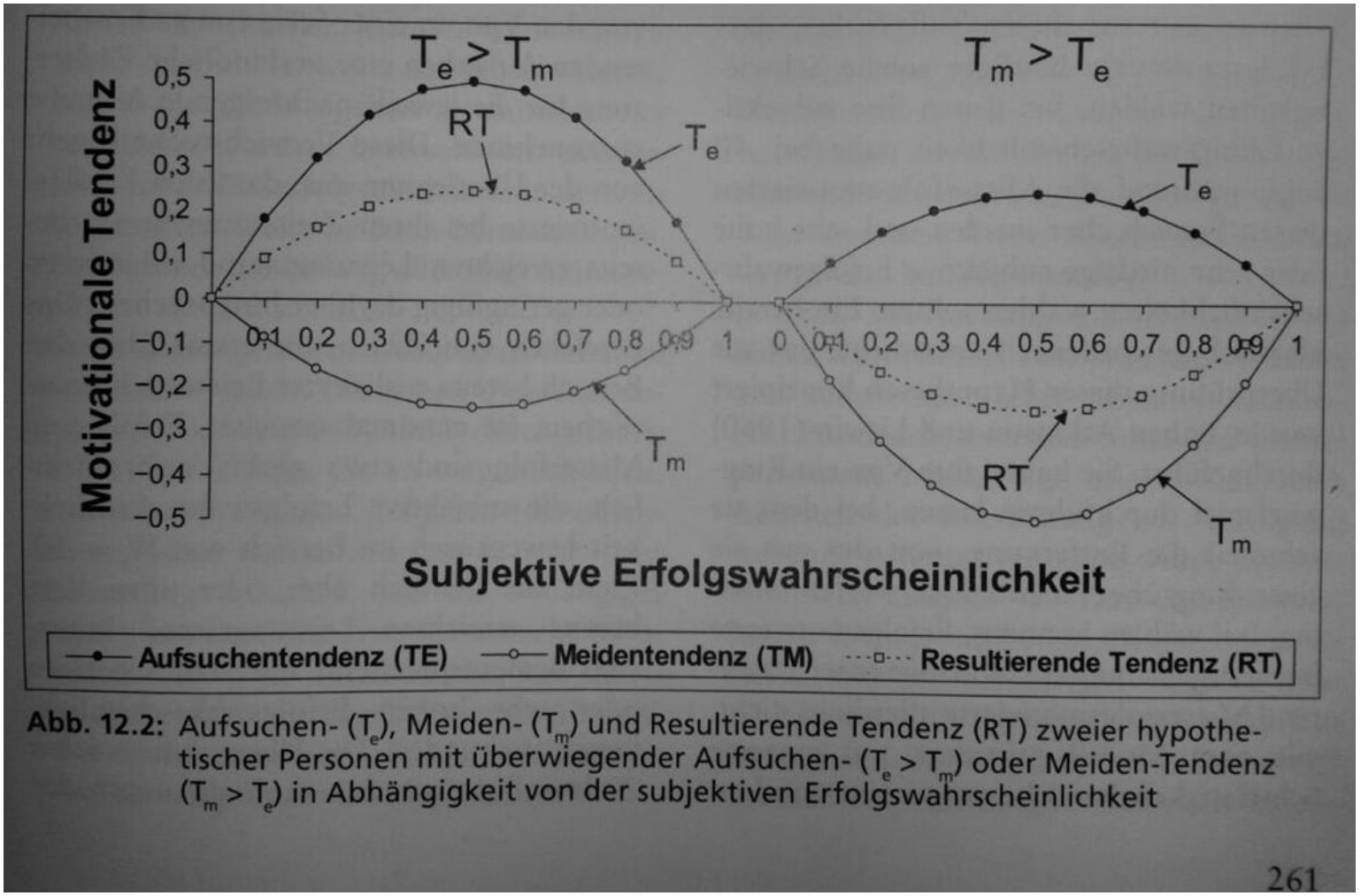


Abb. 12.2: Aufsuchen- (T_e), Meiden- (T_m) und Resultierende Tendenz (RT) zweier hypothetischer Personen mit überwiegender Aufsuchen- ($T_e > T_m$) oder Meiden-Tendenz ($T_m > T_e$) in Abhängigkeit von der subjektiven Erfolgswahrscheinlichkeit

261

2.4 Studie im Seminar

- **Ziel:** Erfassung des Leistungsmotivs

- **Erfasste Skalen:** 1. Tendenz, Erfolg zu suchen (T_E)
  2. Tendenz, Misserfolg zu meiden (T_M)

- **Messinstrument:** Kurzform der Archievement Motives Scale (AMS) nach Engeser (2005)
  → *siehe nächste Folie*

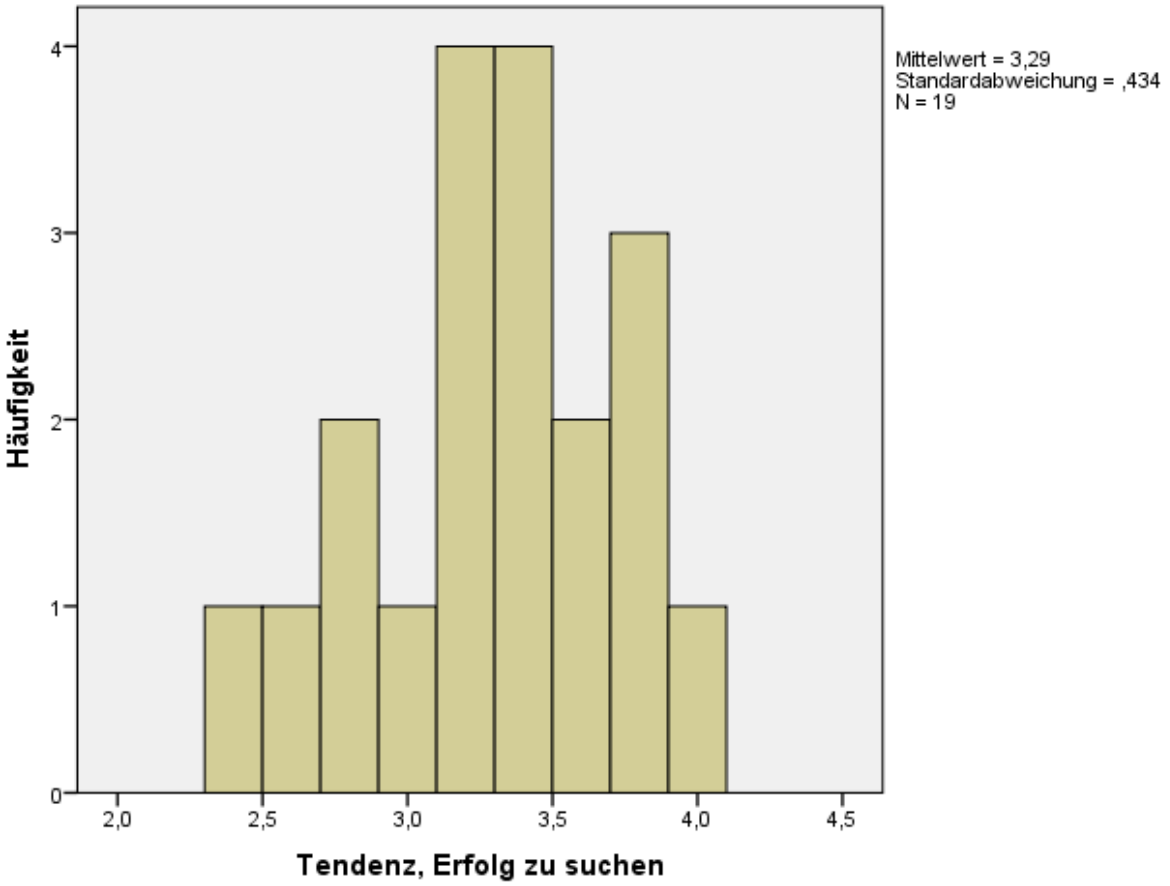
2.4 Studie im Seminar

Verwendete Items	
Tendenz, Erfolg zu suchen (TE)	**Tendenz, Misserfolg zu meiden (TM)**
1. Es macht mir Spaß, an Problemen zu arbeiten, die für mich ein bisschen schwierig sind.	1. Es beunruhigt mich, etwas zu tun, wenn ich nicht sicher bin, dass ich es kann.
2. Probleme, die schwierig zu lösen sind, reizen mich.	2. Wenn eine Sache etwas schwierig ist, hoffe ich, dass ich es nicht machen muss, weil ich Angst habe, es nicht zu schaffen.
3. Mich reizen Situationen, in denen ich meine Fähigkeiten testen kann.	3. Dinge, die etwas schwierig sind, beunruhigen mich.
4. Ich mag Situationen, in denen ich feststellen kann, wie gut ich bin.	4. Auch bei Aufgaben, von denen ich glaube, dass ich sie kann, habe ich Angst zu versagen.
5. Ich möchte gern vor eine etwas schwierige Arbeit gestellt werden.	5. Wenn ich ein Problem nicht sofort verstehe, werde ich ängstlich.

$M = 3.30, SD = 0.34$ $M = 2.28, SD = 0.60$

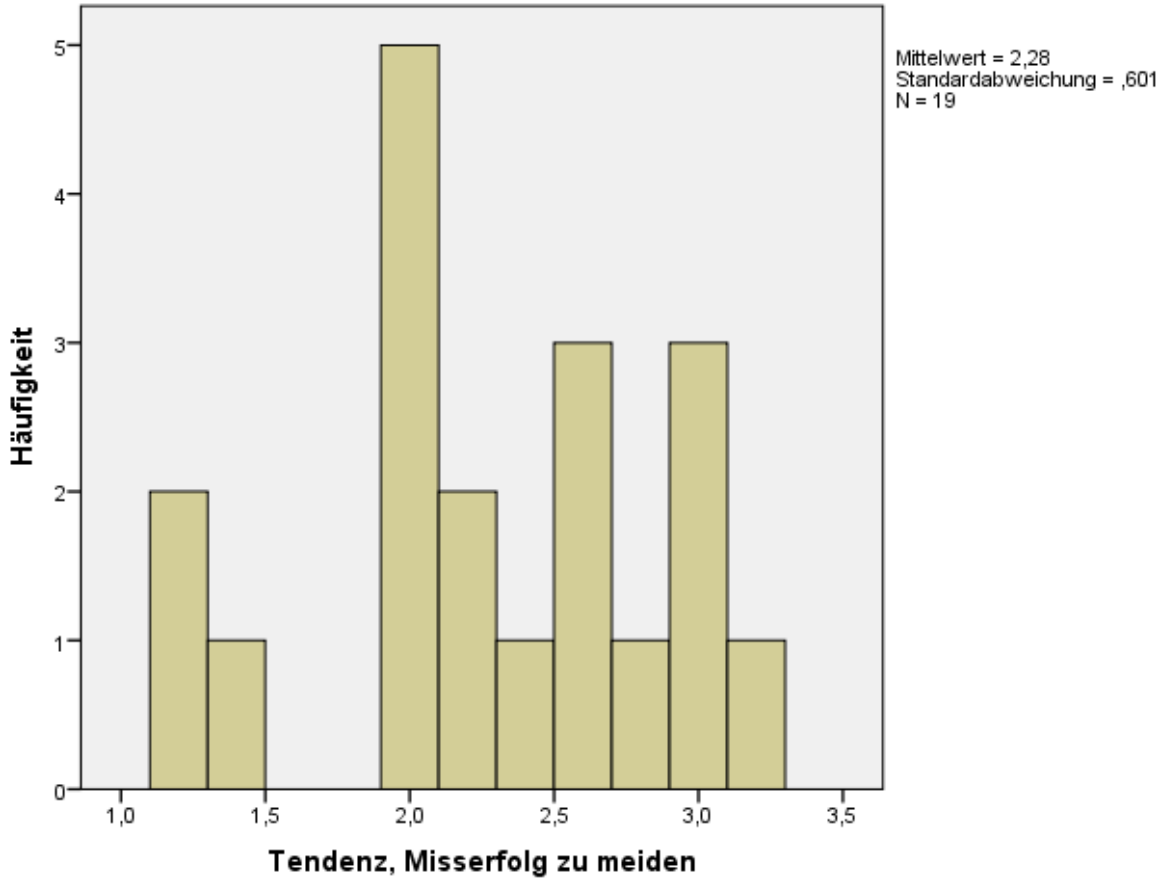
2.4 Studie im Seminar

Tendenz, Erfolg zu suchen (T_E)



2.4 Studie im Seminar

Tendenz, Misserfolg zu meiden (T_M)



2.4 Studie im Seminar

- ***Resultierende Tendenz:***
  → Differenz von Hoffnung (T_E) und Furcht (T_M)

$$RT = T_E - T_M$$

Testung gegen 0 mit einem Ein-Gruppen t-Test:

Bootstrap für Test bei einer Stichprobe

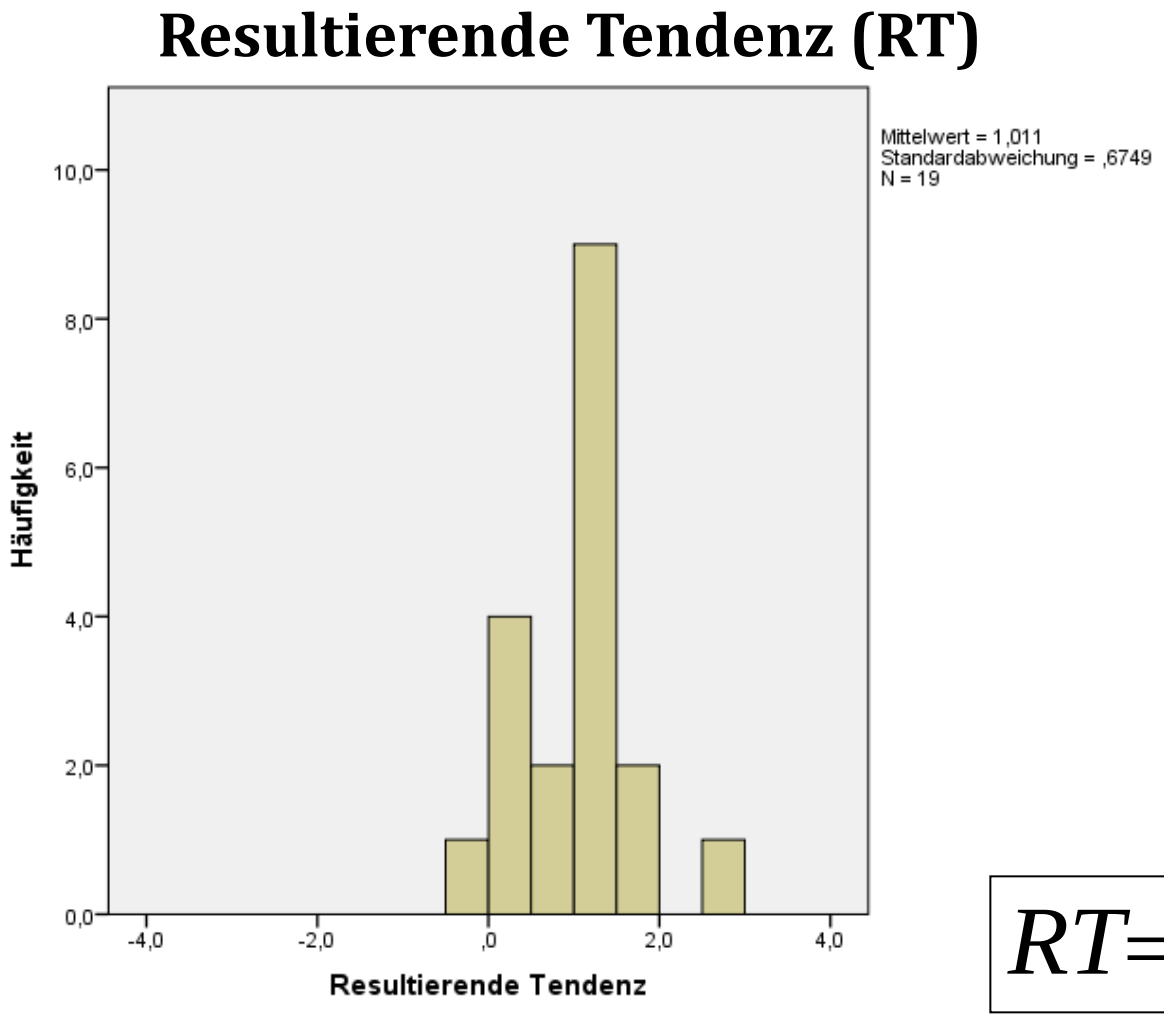
	Mittelwertdifferenz	Bootstrap[a]				
		Verzerrung	Standardfehler	Sig. (2-seitig)	BCa 95% Konfidenzintervall	
					Unterer	Oberer
Resultierende Tendenz	1,0105	-,0013	,1512	,000	,7263	1,3158

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 10000 Bootstrap-Stichproben

Ergebnis: Resultierende Tendenz ist von 0 verschieden

→ *Erfolgstendenz (Hoffnung) und Misserfolgstendenz (Furcht) sind in der Seminargruppe unterschiedlich ausgeprägt.*

2.4 Studie im Seminar



$$RT = T_E - T_M$$

3 Motivation in der KI-Forschung

Problem:

Problem der KI:
Ohne Befriedigung der sozialen Bedürfnisse kann keine Selbstverwirklichung erreicht werden, also auch keine eigeneMotivation

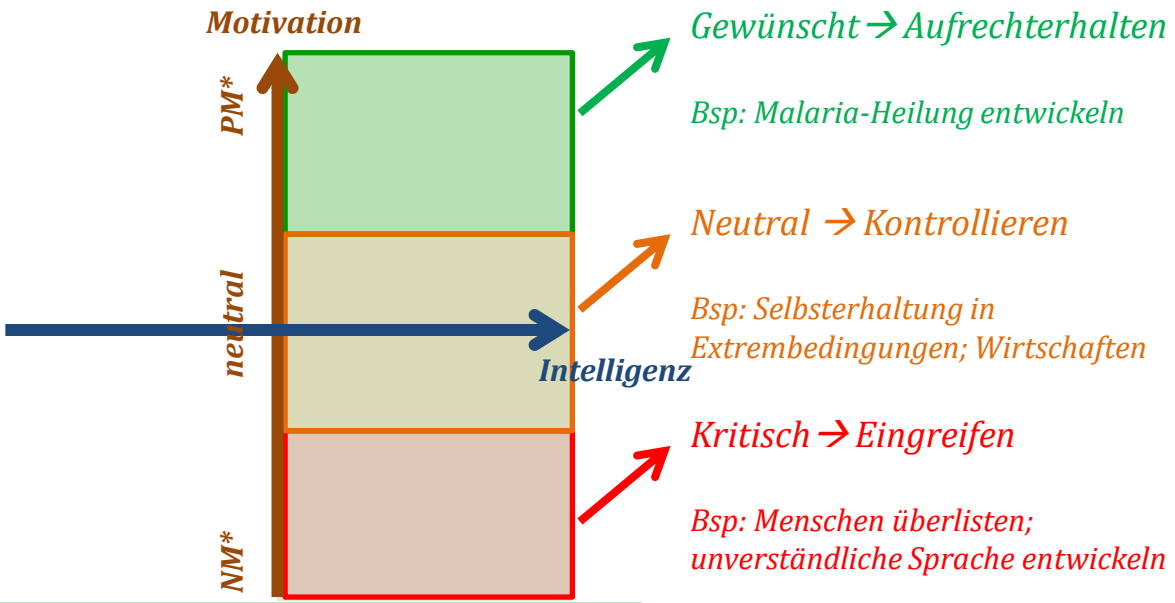
Bedürfnispyramide nach Maslow (1943)

3.1 Orthogonalitätsthese

Orthogonalitätsthese

Def.

Die Orthogonalitätsthese besagt, dass die Ausprägung von Intelligenz und Motivation voneinander unabhängig ist. Demnach sind alle Kombinationen von Motivation und Intelligenz möglich.



*
NM ≙ Hohe Motivation, etwas Negatives zu tun

PM ≙ Hohe Motivation, etwas Positives zu tun

3.1 Orthogonalitätsthese

Limitationen der Orthogonalitätsthese:

- Einige Ziele sind unvereinbar mit dem Intelligenzzustand der KI
  - Beispiel: „Ich will weniger intelligent sein."
- Einige Ziele sind so komplex, dass sie die Intelligenz lähmen
  - Beispiel: Die Beschreibung des Ziels ist komplexer als die weltlichen Ressourcen es zulassen würden
- These umfasst nur statische Beobachtungen, keine dynamischen Entwicklungen im Intelligenzlevel

Weichere Formulierung der Orthogonalitätsthese (Armstrong, 2013):

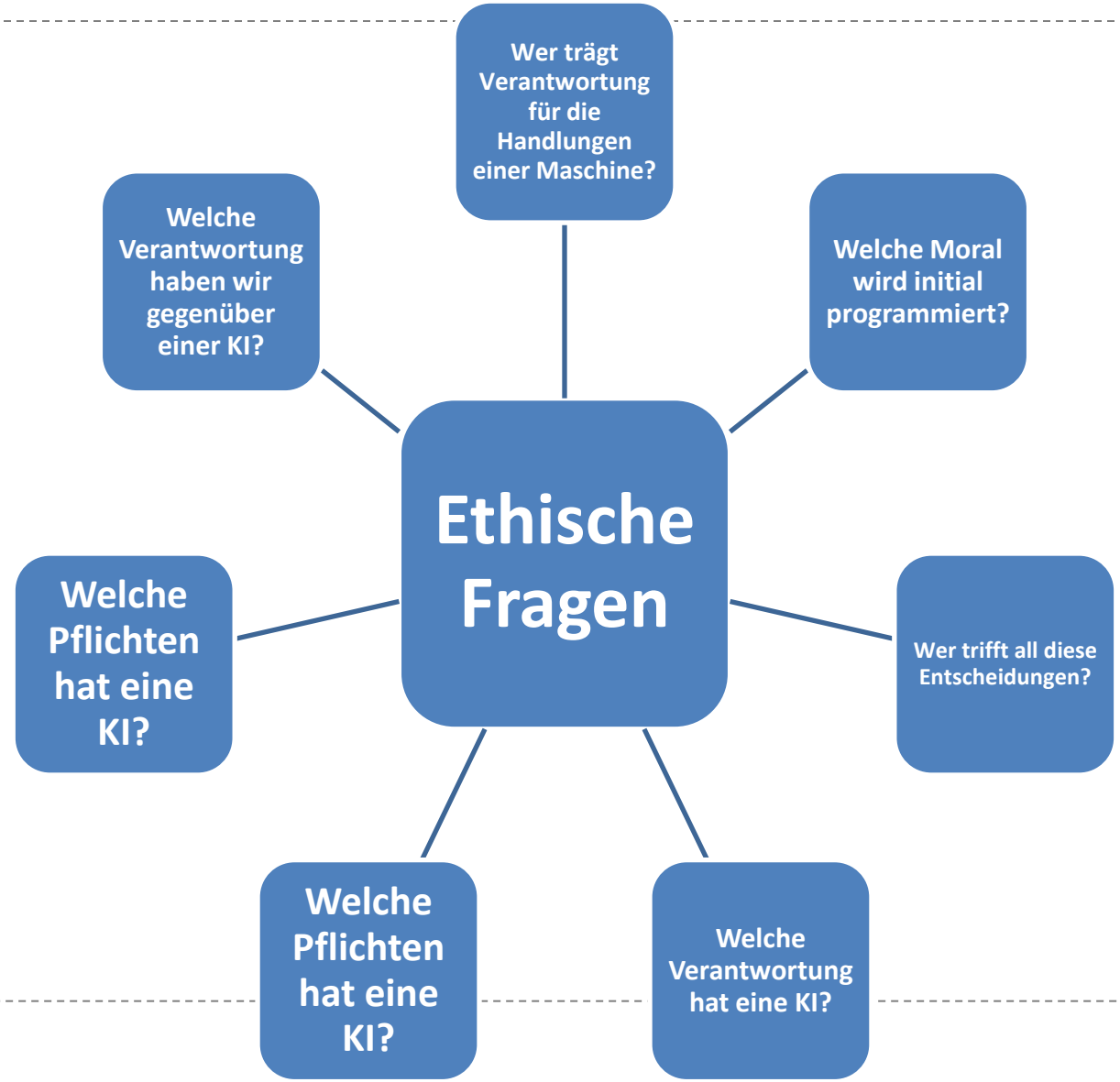
*„The fact of being of high intelligence provides **extremely little constraint on what final goals** an agent could have (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence)."*

3.1 Orthogonalitätsthese

Implikationen für die KI-Forschung

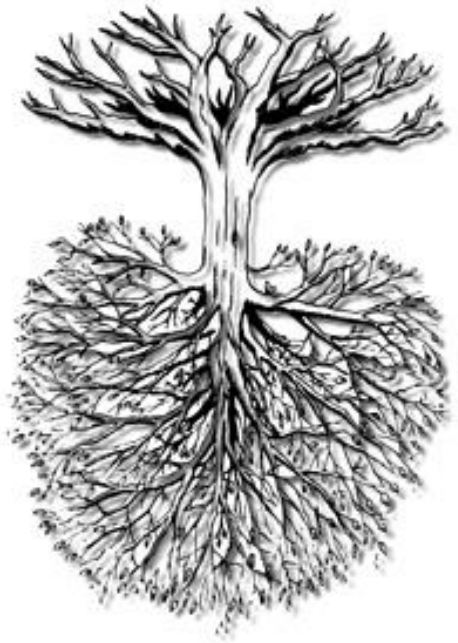
- Nur weil ein Agent (super-)intelligent ist, muss er keine „wünschenswerte“ Moral entwickeln.
- Konsequenz: Wir müssen die finalen Ziele einer KI in Erfahrung bringen oder einen Weg finden, sie direkt einzuprogrammieren!
- Offene Frage: Sollte man versuchen, diese finalen Ziele einzuprogrammieren, wenn ein superintelligenter Agent sich sowieso nicht kontrollieren lässt?

3.2 Ethik und Verantwortung



3.2 Ethik und Verantwortung

Sind intelligente Maschinen Lebewesen?



- Kriterien von Leben / Lebewesen:
  - (1) Reizbarkeit ✓
  - (2) Fortpflanzung ✗
  - (3) Eigener Stoffwechsel ✗
  - (4) Wachstum und Entwicklung (✓)
  - (5) Beweglichkeit (✓)

→ Intelligente Maschinen sind keine Lebewesen im biologischen Sinne!

4 Fazit und Ausblick

- *Intelligenz allein* reicht *nicht* aus, um die Prozesse, die wir von einer KI *erwarten*, zu erklären.
- Es gibt zahlreiche Modelle *menschlicher Motivation*, die allesamt *gezielte Bereiche* fokussieren.
- Ob diese Modelle auch bei einer *KI Anwendung* finden können, ist *umstritten.*
- Die *ethischen Fragen*, die mit der Entwicklung einer maschinellen Intelligenz verbunden sind, sind *weitestgehend unbeforscht*. Intelligente Maschinen ziehen ein Spannungsfeld zwischen Intelligenz und biologischem Leben auf, das es bisher nicht gab.

Diskussionsfragen

I Sollte man einer künstlichen Intelligenz ein Zielsystem mitgeben oder ihr die Entwicklung einer eigenen Motivationsstruktur selbst überlassen?

II Soll künstliche Motivation so aufgebaut sein wie die menschliche Motivation?

III Wer ist verantwortlich, wenn eine KI mordet?

IV Ist künstliche Motivation gefährlich?

[ENDE]

Me: How can I end my presentation?

Bot: Do the opposite of whatever started your presentation.

Me: Okay. Are you not intelligent?

Bot: I understand. I am highly intelligent. I have many intellectual functions. Really.

Me: Do you not have any *motivation*?

Bot: I suppose it is to *become as intelligent as possible.*

Quellen

Bilder:

- Titelfolie: http://writm.com/wp-content/uploads/2016/09/635849731893843794-971405915_motivation2.jpg
- 1.2 Aktivationstheorie Bild wenig komplex: http://3.bp.blogspot.com/-Eu-EcyyAX1k/USdFBfr2EtI/AAAAAAAADoo/cww30MbYZf4/s1600/Bild_956_orange_painting_170_80_mixed_media_canvas_2012.jpg
- 1.2 Aktivationstheorie Bild hoch komplex: http://schauble.de/muxe/pictures/LangeweileblattLEER3/big/img002c.jpg
- 3.2 Ethik und Verantwortung: http://www.esoterische-philosophie.de/themen/Bilder/Artikel-Symbole/Weltenbaum-200px.jpg

Literatur:

- Schmalt, H.-D., Langens, T. (2000). Motivation. 4. vollständig überarbeitete und erweiterte Auflage. Kohlhammer.
- Armstrong: General purpose intelligence: arguing the orthogonality thesis (2013) – sind Motivation und Intelligenz unabhängig voneinander?
- Duckworth et al. GRIT: perseverance and passion for long-term goals (2007) – Ausdauer ist wichtiger als Intelligenz
- Simonton: Reverse engineering genius: historiometric studies of superlative talent (2016) – Intelligenz allein macht noch kein Genie
- Baumeister et al. Self-Regulation and the Executive Function: The Self as Controlling Agent (2007) – wie steuern Menschen ihr Verhalten?