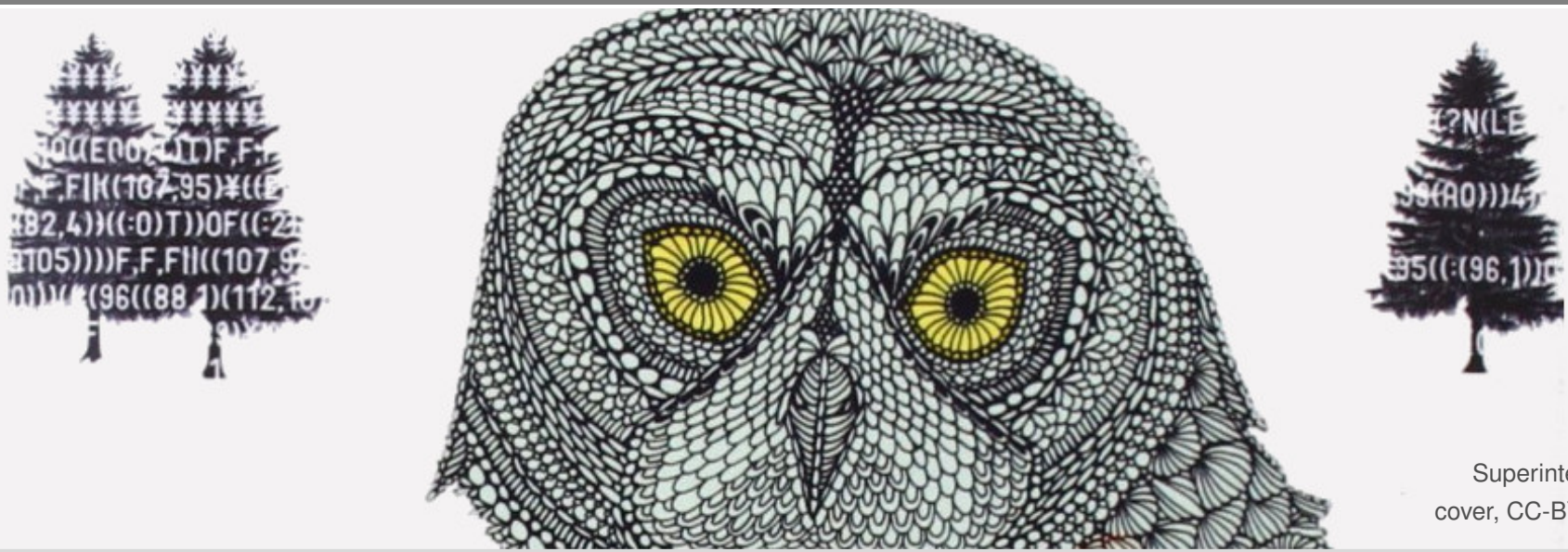
Karlsruher Institut für Technologie

Computer Science meets Philosophy – the Future of AI

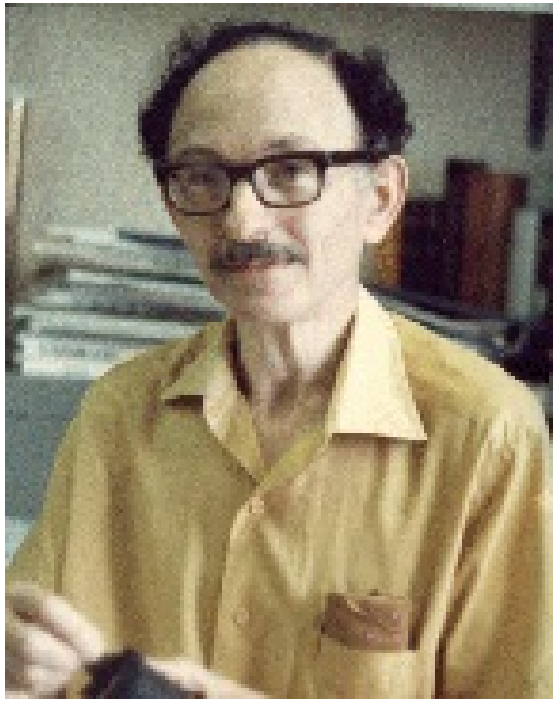
Seminar im Sommersemester 2017
Vorbesprechung, 27.04.2017

INSTITUT FÜR THEORETISCHE INFORMATIK & INSTITUT FÜR PHILOSOPHIE



Superintelligence
cover, CC-BY-SA 4.0

Motivation/Stimmen



“Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.” (I. J. Good, 1965)

https://upload.wikimedia.org/wikipedia/en/b/b4/I._J._Good.jpg

Motivation/Stimmen

Heise: In seinem bekanntesten Werk „Was Computer nicht können“ von 1972 [. . .] zerlegte er die drei Thesen der zeitgenössischen KI-Forschung von Simon: 1. Computer können Schach spielen, 2. Computer können ein mathematisches Theorem lösen und 3. Alle Fragen der menschlichen Psychologie können als Computerprogramm dargestellt werden.

Motivation/Stimmen

“I agree with Elon Musk and some others on this and don’t understand why some people are not concerned.” (Bill Gates)

https://en.wikipedia.org/wiki/File:Bill_Gates_June_2015.jpg, https://en.wikipedia.org/wiki/File:Elon_Musk_2015.jpg, CC-BY-2.0

Zentrale Definitionen nach Bostrom

General intelligence

Possessing common sense and an effective ability to learn, reason and plan to meet complex information-processing challenges across a wide range of natural and abstract domains

Superintelligence

Any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest

Seminar-Anforderungen

- Superintelligenz

Seminar-Anforderungen

- Superintelligenz
  Also natürlich das Buch :-).

Seminar-Anforderungen

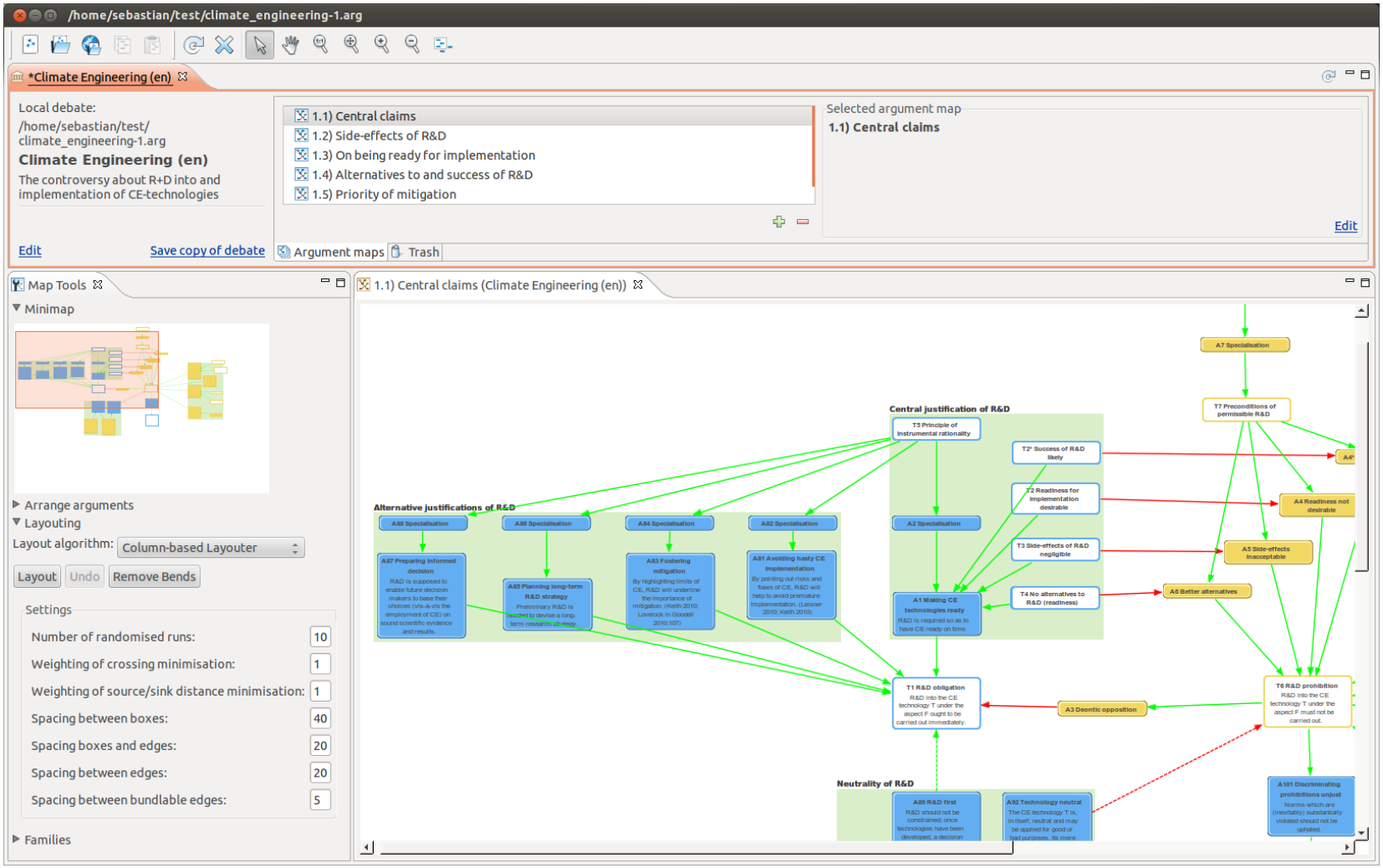
- Superintelligenz
  Also natürlich das Buch :-).
- die Texte der jeweiligen Sitzung lesen & verstehen
- Anwesenheit
- 20min Vortrag & 30min Diskussion leiten
- 3min Kurzpräsentation zu anderen Themen
- Ausarbeitung/Hausarbeit und Review

EUKLID/Schlüsselqualifikation

- Hierbei Argumentanalyse durchführen und vorstellen

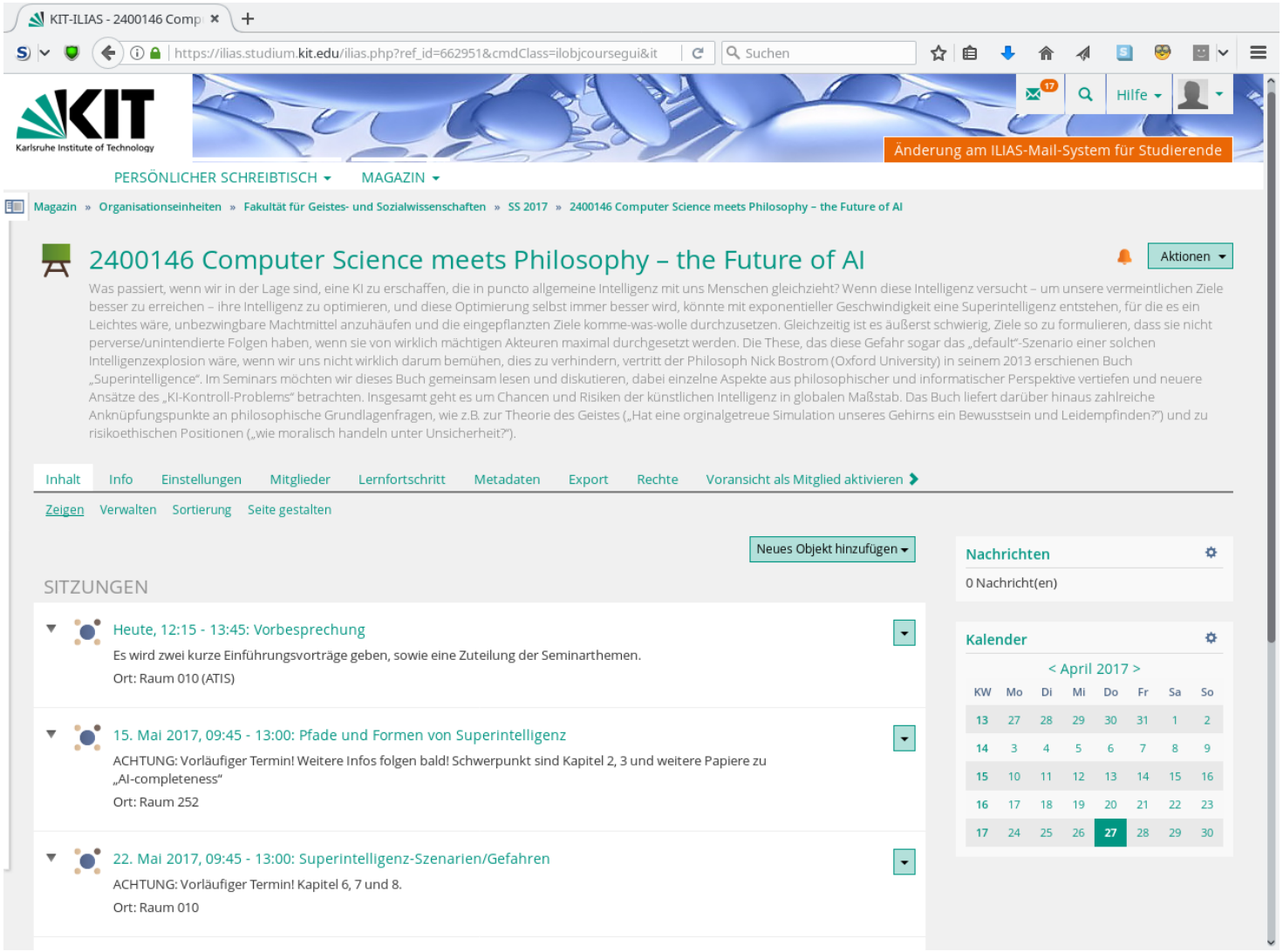
Argumentanalyse

https://en.wikipedia.org/wiki/Argument_map



http://www.argunet.org/wordpress-argunet-2/wp-content/uploads/2013/03/screenshot_01.png

ILIAS


KIT-ILIAS - 2400146 Comp
https://ilias.studium.kit.edu/ilias.php?ref_id=662951&cmdClass=ilobjcoursegui&it
Suchen
Karlsruhe Institute of Technology
Hilfe
Änderung am ILIAS-Mail-System für Studierende
PERSÖNLICHER SCHREIBTISCH
MAGAZIN
Magazin » Organisationseinheiten » Fakultät für Geistes- und Sozialwissenschaften » SS 2017 » 2400146 Computer Science meets Philosophy – the Future of AI
2400146 Computer Science meets Philosophy – the Future of AI
Aktionen
Was passiert, wenn wir in der Lage sind, eine KI zu erschaffen, die in puncto allgemeine Intelligenz mit uns Menschen gleichzieht? Wenn diese Intelligenz versucht – um unsere vermeintlichen Ziele besser zu erreichen – ihre Intelligenz zu optimieren, und diese Optimierung selbst immer besser wird, könnte mit exponentieller Geschwindigkeit eine Superintelligenz entstehen, für die es ein Leichtes wäre, unbezwingbare Machtmittel anzuhäufen und die eingepflanzten Ziele komme-was-wolle durchzusetzen. Gleichzeitig ist es äußerst schwierig, Ziele so zu formulieren, dass sie nicht perverse/unintendierte Folgen haben, wenn sie von wirklich mächtigen Akteuren maximal durchgesetzt werden. Die These, das diese Gefahr sogar das „default"-Szenario einer solchen Intelligenzexplosion wäre, wenn wir uns nicht wirklich darum bemühen, dies zu verhindern, vertritt der Philosoph Nick Bostrom (Oxford University) in seinem 2013 erschienen Buch „Superintelligence". Im Seminars möchten wir dieses Buch gemeinsam lesen und diskutieren, dabei einzelne Aspekte aus philosophischer und informatischer Perspektive vertiefen und neuere Ansätze des „KI-Kontroll-Problems" betrachten. Insgesamt geht es um Chancen und Risiken der künstlichen Intelligenz in globalen Maßstab. Das Buch liefert darüber hinaus zahlreiche Anknüpfungspunkte an philosophische Grundlagenfragen, wie z.B. zur Theorie des Geistes („Hat eine orginalgetreue Simulation unseres Gehirns ein Bewusstsein und Leidempfinden?") und zu risikoethischen Positionen („wie moralisch handeln unter Unsicherheit?").
Inhalt
Info
Einstellungen
Mitglieder
Lernfortschritt
Metadaten
Export
Rechte
Voransicht als Mitglied aktivieren
Zeigen
Verwalten
Sortierung
Seite gestalten
Neues Objekt hinzufügen
SITZUNGEN
Heute, 12:15 - 13:45: Vorbesprechung
Es wird zwei kurze Einführungsvorträge geben, sowie eine Zuteilung der Seminarthemen.
Ort: Raum 010 (ATIS)
15. Mai 2017, 09:45 - 13:00: Pfade und Formen von Superintelligenz
ACHTUNG: Vorläufiger Termin! Weitere Infos folgen bald! Schwerpunkt sind Kapitel 2, 3 und weitere Papiere zu „AI-completeness"
Ort: Raum 252
22. Mai 2017, 09:45 - 13:00: Superintelligenz-Szenarien/Gefahren
ACHTUNG: Vorläufiger Termin! Kapitel 6, 7 und 8.
Ort: Raum 010
Nachrichten
0 Nachricht(en)
Kalender
< April 2017 >
KW Mo Di Mi Do Fr Sa So
13 27 28 29 30 31 1 2
14 3 4 5 6 7 8 9
15 10 11 12 13 14 15 16
16 17 18 19 20 21 22 23
17 24 25 26 27 28 29 30

Block 1: Pfade und Formen von Superintelligenz

Mo, 15. Mai, 9:45 – 13:00 Uhr

1./2. (Inf.) Welche Wege könnte es hin zu einer Superintelligenz geben:

1. Whole Brain Emulation
2. Biologische Wege
3. Bessere ML-Algorithmen
4. . . .

2. (Inf.) Welche Formen:

1. „Speed superintelligence“
2. „Quality superintelligence“

3. Welche komplexitätstheoretische Fragen sind relevant? (P = NP? Was heißt AI-complete?) (Inf.)

Buchstelle: Kapitel 2 und 3, Quellen zu komplexitätstheoretischen Fragen

Block 2: Superintelligenz-Szenarien/Gefahren

Mo, 22. Mai, 9:45 – 13:00 Uhr

- 4. Cognitive Superpowers / AI takeover scenarios (Inf.)
- 5. Zusammenhang Intelligenz – Ziele/Motivation (SQ oder Inf.)
- 6. Sind Auslöschungszenarien/Dystopien der „default“? (SQ)

Buchstelle: Kapitel 6, 7 und 8.

Block 3: KI-Kontroll-Problem & Lösungsansätze

Mo, 29. Mai, 9:45 – 13:00 Uhr

- 7. Fähigkeitsbeschränkungen zur KI-Kontrolle (Inf.)
- 8. Motivationsauswahl zur KI-Kontrolle (Inf.)
- 9. i) Mind Crime, „Können (Gehirn-)Simulationen Leid empfinden?“ (SQ)
  ii) Risikoethik, Formalisierung/von Neumann–Morgenstern Axiome (Inf.)

Buchstelle: Kapitel 8, 9 und 10.

Block 4: Wie bringen wir KI's Werte bei?

Fr, 2. Juni, 9:45 – 11:15 Uhr, 14:00 – 15:30 Uhr

- 10. Ansätze um KIs Werte beizubringen (Inf. oder SQ)
- 11. Axiomatisierung von Werteauswahlkriterien, Entscheidungsverfahren (Inf.)
- 12. Analyse der coherent extrapolated volition (SQ)

Buchstelle: Kapitel 12 und 13.

Block 5: Strategien, Technikfolgenabschätzung

Mo, 12. Juni, 9:45 – 13:00 Uhr

- 13. Kartierung möglicher Strategien (SQ)
- 14. Technikfolgenabschätzung (SQ)

Abschlussdiskussion

Buchstelle: Kapitel 14.