Was können wir aus der Entwicklungspsychologie lernen, um Interaktionen mit Chat-Bots, die auf LLMs basieren, zu gestalten und zu verstehen?

DENKWERKSTATT BERLIN

Anna Strasser (2025)

Slides are downloadable at

Slides can be downloaded at https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS

Generative KI – eine kontroverse Diskussion

ARTIFICIAL INTELLIGENCE | MAR. 1, 2023

You Are Not a Parrot

And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this.

By Elizabeth Weil, a features writer at New York

OPINION

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

By Gary Marcus & Ernest Davis

August 22, 2020

OpenAI

February 24, 2023

Planning for AGI and beyond

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

Blake Lemoine Follow

Jun 11 · 20 min read · Listen

Save

Is LaMDA Sentient? — an Interview

What follows is the "interview" I and a collaborator at Google conducted with LaMDA. Due to technical limitations the interview was conducted over several distinct chat sessions. We edited those sections together into a single whole and where edits were necessary for readability we edited our prompts but never LaMDA's responses. Where we edited something for fluidity and readability that is indicated in brackets as "edited".

Taking AI Welfare Seriously

Robert Long* Eleos AI

Jeff Sebo* New York University

Patrick Butlin† University of Oxford

Kathleen Finlinson† Eleos AI

Kyle Fish†§ Eleos AI, Anthropic

Jacqueline Harding† Stanford University

Jacob Pfau† New York University

Toni Sims† New York University

Jonathan Birch‡ London School of Economics

David Chalmers‡ New York University

Anthropic has hired an 'AI welfare' researcher

Kyle Fish joined the company last month to explore whether we might have moral obligations to AI systems

(Bender et al., 2021; Hashim, 2024; Lemoine, 2022; Long et al., 2024; Marcus & Davis, 2020; Open-AI, n.d.; Weil, 2023)

Ein begriffliches Problem

Ist ein LLM-basierter Chatbot ein soziales Gegenüber oder nur ein Ding?

HARD-CORE INSTRUMENTAL VIEW
nur ein unbelebtes Ding

IN EXPECTATION OF AGI VIEW
ein mögliches soziales Gegenüber

INBETWEENISM
weder noch
-
irgendetwas zwischen echten Persönlichkeiten & rein kausal beschreibbaren Maschinen

➢ **WIR KÖNNEN NICHT ALLE UNSERE INTERAKTIONEN MIT LLMS AUF WERKZEUGGEBRAUCH REDUZIEREN**

… it is neither quite right to say that our interactions with LLMs are properly asocial (just tool-use or self-talk) nor quite right to say that our interactions with LLMs are properly social. Neither standard philosophical theorizing nor dichotomous ordinary concepts enable us to think well about these in-between phenomena.
(Strasser & Schwitzgebel 2024, *p. 197)*

Warum Entwicklungspsychologie?

UNTERSCHIEDLICHE TYPEN VON ASYMMETRISCHEN INTERAKTIONEN

ENTWICKLUNGSPSYCHOLOGIE

- gemeinsames Handeln von Erwachsenen und Kindern
- Kinder = **sozial interagierende Wesen**

KÜNSTLICHE INTELLIGENZ

- gemeinsames Handeln von Menschen und künstlichen Systemen
- künstliche Systeme **=?=** **sozial interagierende Entitäten**

ERWACHSENE & KINDER

MENSCHEN & KÜNSTLICHE SYSTEME

KEINE NOTWENDIGKEIT, DASS DIE FÄHIGKEITEN GLEICHMÄßIG AUF ALLE TEILNEHMER VERTEILT SIND

Um Missverständnisse zu vermeiden, möchte ich betonen, dass ich Interaktionen mit Kindern nicht mit Interaktionen mit künstlichen Systemen gleichsetze – sie haben nur die Eigenschaft gemeinsam, dass sie asymmetrisch sind.

Graduelle Ansätze & minimale Begriffe

Philosophische Begriffe stellen zu anspruchsvolle Anforderungen

Philosophen neigen dazu, Idealfälle zu beschreiben → Kinder, nicht-menschliche Tiere und Roboter (künstliche Agenten) fallen oft durch das begriffliche Netz

Artist: Lorin Strasser

→ MINIMAL APPROACHES

- wie man die anspruchsvolle Terminologie der Philosophie erweitern kann, um Phänomene zu erfassen, die man in der Entwicklungspsychologie, der Tierkognition und der KI vorfindet

Stephen Butterfill & Ian Apperly (2013): minimal mindreading | John Michael et al. (2016): minimal sense of Commitment | Elisabeth Pacherie (2013): shared intention lite
Anna Strasser (2006): minimal action

Ein multidimensionales Spektrum sozialer Interaktionen

INBETWEEN PHENOMÄNE

QUASI-SOZIALE ASYMMETRISCHE INTERAKTIONEN

keine soziale Interaktion

einseitige Sozialität
- Werkzeug-gebrauch

quasi-social human-machine interaction

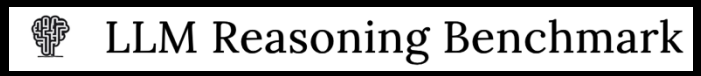
quasi-social human-animal interaction

quasi-social adult-infant interaction

social adult-adult interaction

Vollwertige, intellektuell anspruchsvolle, kooperative soziale Interaktion

Philosophische Begriffe sind Gegenstand der Diskussion



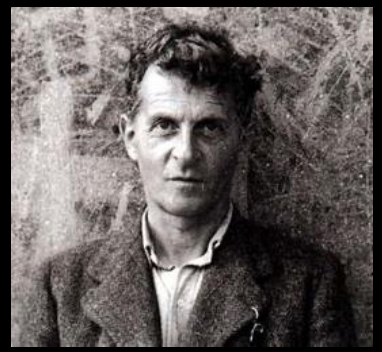
Viele Begriffe, die von Philosophen bisher der Beschreibung der besonderen Merkmale des Menschen als rationalem Akteur vorbehalten waren, werden nun auf Maschinen angewandt.

Das führt zu intensiven Debatten über Begriffe wie Verstehen, Wissen, Argumentation und phänomenologisches Bewusstsein.

(Agrawal et al., 2023; y Arcas, 2022; Salonen, n.d.; Trott et al., 2023; Lake & Baroni, 2023)

Familienähnlichkeit & multiple Realisierungen

- Vielleicht brauchen Maschinen gar kein Bewusstsein, um sich als quasi-soziale Interaktionspartner zu qualifizieren?
- Vielleicht gibt es minimale Formen von Handlungsfähigkeit?
- Vielleicht ist es auch denkbar, dass es multiple Realisationen von manchen sozio-kognitiven Fähigkeiten gibt?



Wittgenstein, Ludwig (2003). *Philosophische Untersuchungen*.

ALLE INSTANZEN STEHEN IN EINEM VERHÄLTNIS VON
FAMILIENÄHNLICHKEIT & ZULÄSSIGKEIT VON MULTIPLEN REALISIERUNGEN
VON BEDINGUNGEN

- CONTRA ganzes Paket von anspruchsvollen Bedingungen, die notwendigerweise gleichzeitig auftreten müssen
- PRO verschiedene Kombinationen von Bedingungen & unterschiedliche Ausprägungen der Bedingungen

DISJUNKTIVER BEGRIFFSRAHMEN

- der sowohl minimale Erscheinungsformen von Bedingungen als auch multiple Realisationen zulässt

KANN MENSCH-MASCHINEN-INTERAKTIONEN UND DIE FÄHIGKEITEN DER KÜNSTLICHEN SYSTEME ANGEMESSEN CHARAKTERISIEREN

Kommunikation & Verstehen

Es ist fraglich, ob LLMs selbst verstehen können, was ihre Outputs für uns bedeuten.

ARGUMENTE GEGEN DIE ZUSCHREIBUNG VON VERSTEHEN

LLMs

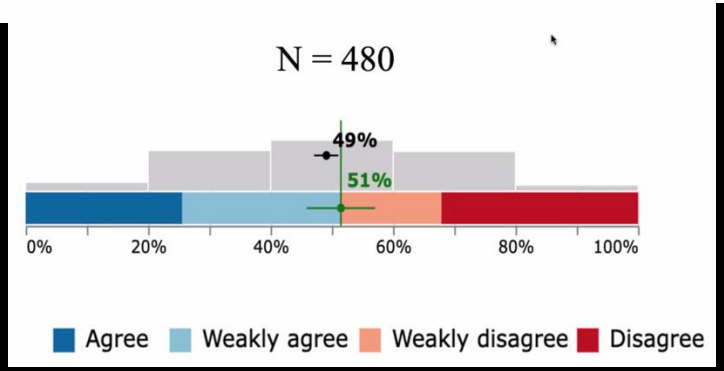
- sind nicht in der Lage, eine Welt mit uns zu teilen
- sind nicht ge*grounded*
- haben keine kommunikativen Intentionen
- sind nicht darauf trainiert, die Wahrheit von Äußerungen zu berücksichtigen
  - Outputs von LLMs sind nicht zuverlässig, sie halluzinieren, sie machen schwere Fehler ...

WHAT DO NLP RESEARCHERS BELIEVE?
RESULTS OF THE NLP COMMUNITY METASURVEY

2022

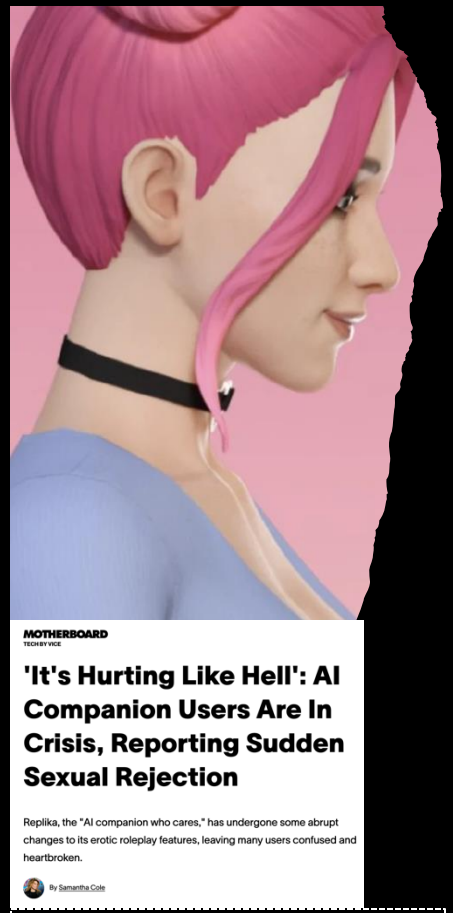
Julian Michael[1,2] Ari Holtzman[1] Alicia Parrish[4] Aaron Mueller[5] Alex Wang[3]
Angelica Chen[2] Divyam Madaan[3] Nikita Nangia[2]
Richard Yuanzhe Pang[3] Jason Phang[2] and
Samuel R. Bowman[2,3,4]

Agree or disagree: Some generative models trained only on text, given enough data and computational resources, could understand natural language in some non-trivial sense.



Michael et al., 2022

Eklatante Ähnlichkeiten mit Kommunikation



MOTHERBOARD
TECH BY VICE

'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection

Replika, the "AI companion who cares," has undergone some abrupt changes to its erotic roleplay features, leaving many users confused and heartbroken.

By Samantha Cole

- 2023

Replika users feel like losing their best friend after an update

Menschen interagieren jedoch mit LLMs auf eine Art und Weise, die echten Gesprächen mit sozialen Beziehungspartnern stark ähneln.

Wir sind zu Recht verwirrt darüber, welche Fähigkeiten LLMs wirklich haben und welche sie nur scheinbar haben.

Aus meiner Sicht ist es wichtig, dass wir sowohl die Ähnlichkeiten als auch die Unterschiede herausarbeiten.

→ asymmetrischen Merkmale von Mensch-Maschinen Interaktionen

Kommunikation ohne vollausgeprägtes Verstehen?

Sind LLMs von sozialen Interaktionen, wie Kommunikation, ausgeschlossen, wenn wir davon überzeugt sind, dass sie sprachliche Äußerungen nicht so verstehen (begreifen) können wie wir Menschen?

1. genaueren Betrachtung der Unterschiede zwischen dem, was Daniel Dennett als *competence with comprehension* und *competence without comprehension* beschrieben hat

2. Gibt es Beispiele für die Kompetenz zu ein gegenüber in einer Kommunikation zu sein, ohne dass man von ‚echtem' Verstehen sprechen muss?

3. **kommunikative Situationen, in denen es offensichtlich ist, dass einer der Beteiligten wenig oder vielleicht nicht wirklich versteht worüber gesprochen wird**
   - **kommunikative Situationen mit Kindern**

kommunikative Aktivitäten oder Sprachspiele können in der Verteilung der Fähigkeiten asymmetrisch sein

Inwieweit kann die Analyse solcher Sprachspiele eine hilfreiche Vorlage für die Beschreibung menschlicher Interaktionen mit LLMs bieten?

Verstehen oder Nicht-Verstehen?

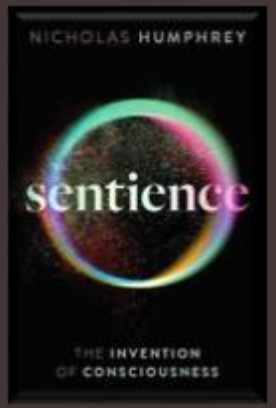
Können LLMs einen paradigmatischen Fall von nicht-lebenden Entitäten darstellen, denen man eine neue Art des Verstehens zuschreiben kann – eine minimale Art des Verstehens, das nicht alle Merkmale des menschlichen Verstehens aufweist, insbesondere nicht die Merkmale des Bewusstseins oder der Empfindungsfähigkeit.

Ich denke momentan nicht mehr, dass wir gute Gründe haben, den LLMs Verstehen zuzuschreiben!
Nun frage ich mich, ob LLMs überhaupt verstehen müssen, um als Kommunikationspartner zu fungieren.

In den Debatten über mögliche Zuschreibungen von Bedingungen wird eine Alles-oder-Nichts-Haltung eingenommen. Zuschreibungen werden so behandelt, als ob sie immer im Paket kämen.

ABER: Nicht alle Dinge kommen in einem Paket!

- Dennett:
  Pflanzen & Bakterien sind empfindungsfähig, aber nicht bewusst
- Humphrey:
  Entitäten mit kognitivem Bewusstsein, aber ohne phänomenales Bewusstsein



(Cukier, 2022; Humphrey, 2023)

Dennett & the four creatures

	Implementation	Eigenschaften	Verstehen?	Lernen?
Gregorian 	bewusstes Einsetzen der Denkwerkzeuge	*können Gelerntes auf Neues anwenden*	JA verstehen die Gründe für ihr eigenes Verständnis	JA!!!
Popperian 	Maxime: antizipieren bevor sie handeln → haben eine Vorliebe für Vorabprüfung (*pretesting*)	*informationssensible & zukunftsorientierte Verarbeitungs-prozesse*	NEIN verstehen nicht, warum sie antizipieren	JA !! Antizipieren von Konsequenzen der Handlungsoptionen basierend auf gespeicherten Information über die Welt
Skinnerian 	fest verdrahtet bevorzugen, was auch immer verstärkende Effekte hat	*gewisse Plastizität des Verhaltensrepertoire*	NEIN verstehen nicht, warum sie aus Versuch und Irrtum lernen	JA Versuch und Irrtum
Darwinian 	fest verdrahtet	*ahnungslos gegenüber neuen Situationen*	NEIN nur angeborene Fähigkeiten	NEIN

(Dennett, 2018)

competence with comprehension

NUR VOLLWERTIGE (FULL-FLEDGED) AUTONOME AGENTEN !

Finally, could you turn LaMDA or GPT-3 into a language-user, capable of speech acts, and if so, how?
By allowing it to cherry-pick its own outputs. But for it to do this, it must be an agent with a high degree of autonomy, capable of revising its own cherry-picking processes to better meet its goals.
(Dennett 2024, p.28)

Da ich davon überzeugt bin, dass künstliche Systeme noch nicht als vollwertige autonome Agenten gelten können, auch wenn ich ihnen minimale Handlungsfähigkeit in quasi-sozialen Interaktionen mit Menschen zuschreiben würde, bin ich motiviert zu untersuchen, ob alle kommunikativen Situationen das Verstehen aller Beteiligten voraussetzen.

Verstehen

KEINE ALLGEMEIN ANERKANNTE THEORIE DES VERSTEHENS

Folgt man Dennett mit seinen Ausführungen zu den *four creatures*
→ sehr anspruchsvollen Begriff von Verstehen

- Aber Dennett wäre nicht Dennett wenn er nicht gleichzeitig für einen graduellen Ansatz argumentieren würde und dabei findet man Stellen, in denen er von *sorta comprehension* spricht.

Das Vorliegen von Verstehen sollte man nicht als eine Alles- oder Nichts Frage behandeln.

INBETWEEN

Denken Sie an alle Arten von sprachlichem Austausch, den Sie in Ihrem Leben hatten

Keine Kommunikation

- aneinander vorbeireden
- nichts weiter als zwei Einheiten, die sich beim Sprechen abwechseln

Erfolgreiche Kommunikation

- echter Austausch, bei dem sich beide Interaktionspartner verstehen

Fälle, die in keine der beiden Kategorien fallen

– INBETWEEN PHÄNOMENE –

- mehr oder weniger erfolgreiche Kommunikation
- mit *sorta comprehension*
- mit einem asymmetrischen Aspekt

Mit Kindern sprechen

Der Übergang von einer Kompetenz mit *sorta comprehension* zu einer Kompetenz mit vollausgeprägtem Verstehen ist ein gradueller

Komplexe soziale Fähigkeiten entstehen nicht von einem Moment auf den anderen

- nicht in der Entwicklung des Menschen,
- nicht in der Phylogenese der Tiere und
- nicht technologisch im Design von KI-Systemen.

- Kinder verwenden häufig Wörter, ohne sie zu verstehen
  - sie wiederholen einfach etwas, das sie zuvor gehört haben

ABER

- durch wiederholte Interaktionsabfolgen, bei denen sie die Reaktionen ihrer Interaktionspartner beobachten können, beginnen sie, die Bedeutung immer besser zu verstehen

sprachliche Interaktionen mit Kindern sollte man klar als Kommunikation kategorisieren

gradueller Ansatz:

➢ Kommunikation in ihren Anfangsphasen kann als asymmetrische* Interaktion beschrieben werden

**asymmetrisch*:*

Die Bedingungen, die von den Interaktionspartnern erfüllt werden müssen, unterscheiden sich zwischen Kindern und Erwachsenen

Unterschiede

EIN MENSCHLICHES FEATURE

Indem wir mit Kindern so sprechen, als würden sie alles verstehen, geben wir ihnen die Chance, immer mehr zu verstehen.

ABER

die derzeitigen KI-Systeme sind nicht in der Lage, aus unseren Interaktionen mit ihnen zu lernen, so wie es Kinder tun

.

Wenn dies in Zukunft für Maschinen zutreffen sollte, dann könnten wir sagen, dass die Behandlung als soziale Partner ihnen dabei helfen könnte, die Reaktionsmuster so weiterzuentwickeln, die sie zu sozialeren Partnern machen.

Wir können aber auf Fälle von Kommunikation in unserer täglichen Erfahrung verweisen, in denen dem Kommunikationspartner ein großes Maß an Verstehen fehlt.

Weitere Beispiele

Mit sehr betrunkenen Menschen sprechen

Studierende in Prüfungssituationen

Bullshitting

Small talk

Nicht alle Sprachspiele setzen ein vollausgeprägtes Verstehen voraus.

Manchmal scheint es ausreichend zu sein, ein paar „einfache“ Regeln zu befolgen, Muster zu wiederholen, die wir zuvor beobachtet haben, und alle möglichen Chat-Spiele zu spielen.

Keine Konklusion – sondern viele Fragen

ES GIBT

KOMMUNIKATIVE AKTIVITÄTEN ODER SPRACHSPIELE,

BEI DENEN DIE FÄHIGKEITEN UNGLEICH VERTEILT SIND.

Sind LLMs

- wie Kinder, die nie erwachsen werden und nichts dazu lernen
- wie betrunkene Kommunikationspartner, die nie nüchtern werden
- wie geschickte Sprachspieler, die nur über *competence without comprehension* verfügen?

All dies wäre nicht möglich gewesen, wenn ich nicht die Möglichkeit gehabt hätte mit vielen Menschen und Maschinen zu interagieren.

Daniel Dennett

Eric Schwitzgebel

Joshua Rust

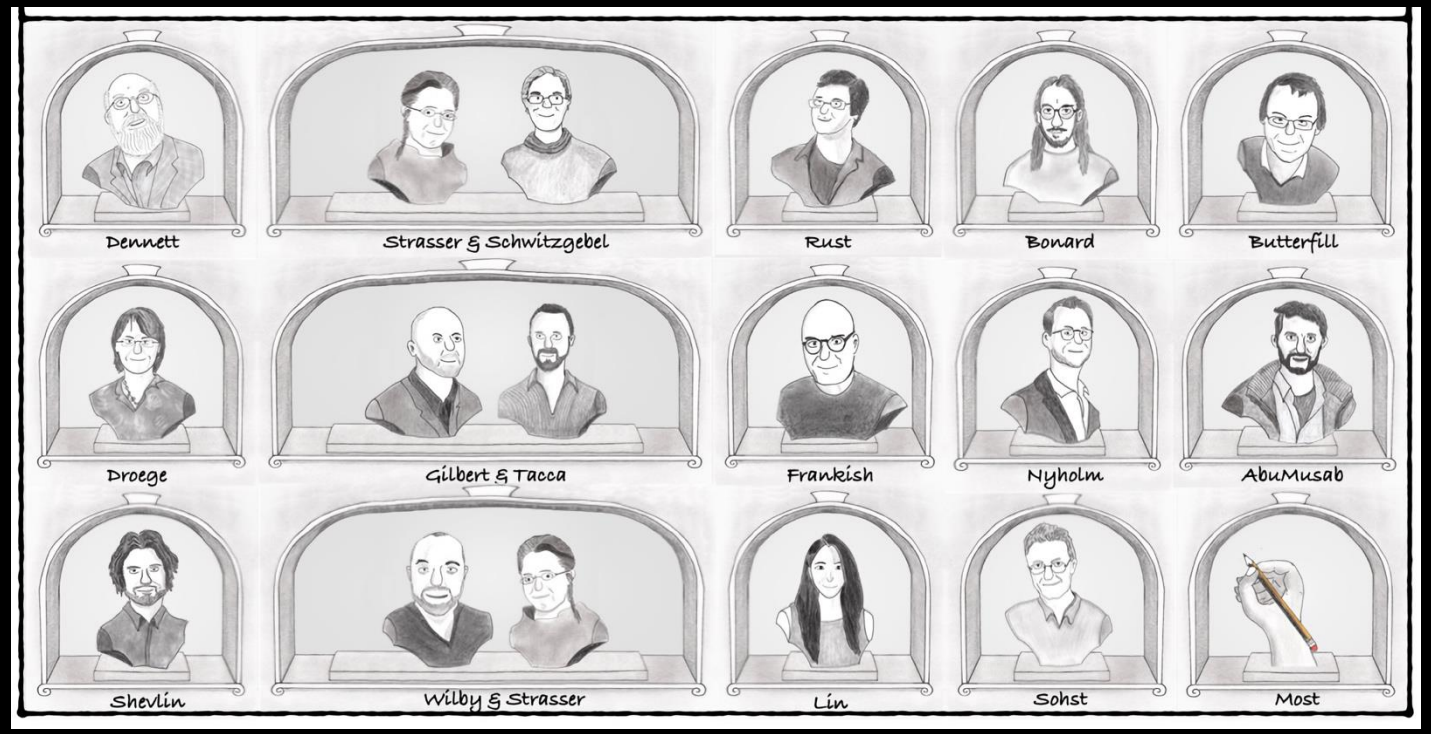
Steven Butterfill

Mike Wilby

DigiDan

Vielen Dank fürs zuhören!

Werbung in eigener Sache



Referenzen

Agrawal, A., Mackey, L., & Kalai, A. T. (2023). *Do Language Models Know When They're Hallucinating References?* (arXiv:2305.18248). arXiv. http://arxiv.org/abs/2305.18248

Agüera y Arcas, B. (2022). Do Large Language Models Understand Us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Butterfill, S., & Apperly, I. (2013). How to Construct a Minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Hashim, S. (2024, December 5). *Anthropic has hired an "AI welfare" researcher*. https://www.transformernews.ai/p/anthropic-ai-welfare-researcher

Lake, B. M., & Baroni, M. (2023). Human-like Systematic Generalization through a Meta-learning Neural Network. *Nature*, 1–7. https://doi.org/10.1038/s41586-023-06668-3

Lemoine, B. (2022, June 11). Is LaMDA Sentient? — An Interview. *Medium*. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI Welfare Seriously* (arXiv:2411.00986). arXiv. https://doi.org/10.48550/arXiv.2411.00986

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview. com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2022). *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey* (arXiv:2208.12852). arXiv. https://doi.org/10.48550/arXiv.2208.12852

Michael, J., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01968

Open-AI. (n.d.). *Planning for AGI and beyond*. Retrieved May 6, 2024, from https://openai.com/index/planning-for-agi-and-beyond

Pacherie, E. (2013). Intentional joint agency: Shared intention lite. *Synthese*, *190*(10), 1817–1839. https://doi.org/10.1007/s11229-013-0263-7

Salonen, S. (n.d.). *LLM Reasoning Benchmark*. Retrieved February 17, 2025, from https://www.llm-reasoning-benchmark.com/

Strasser, A. (2006). Kognition künstlicher Systeme. In *Kognition künstlicher Systeme*. De Gruyter. https://doi.org/10.1515/9783110321104

Strasser, A., & Schwitzgebel, E. (2024). Quasi-sociality: Toward Asymmetric Joint Actions. In *Anna's AI Anthology. How to live with smart machines?* xenomoi Verlag.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111/cogs.13309

Weil, E. (2023, March 1). *You Are Not a Parrot*. Intelligencer. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

Wittgenstein, L. (2003). *Philosophische Untersuchungen* (J. Schulte, Ed.). Suhrkamp.