AI 507: Artificial Intelligence and Society

#3 Social cognition

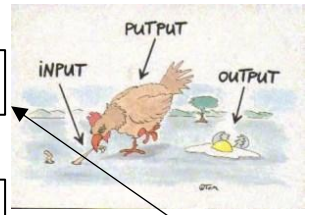
Associate Prof. Dr. Lena Frischlich

Flashlight

SDU DDC
digital democracy centre

Flashlight

Memory



- Encoding
- Storage
  - Short-term → 4-7 units
  - Long-term
- Retrieval
  - Cues ease retrieval

AI & Learning

- Negative effects
- Positive effects
- Context matters!
- Dunning-Kruger effect

Schemes

Learning

- Temporal contiguity →
- Predictor selection
- Configurations
- Generalisation
- Un-learning

Classic behaviourism

Perception

- Attentional blindness
- Perception = construction

Practical consequences

Any questions to last week?

Then, let’s get started!

After today’s lecture, I hope you will…

- … have insights into social cognition
- …understand how social groups shape how we think
- … have a better understanding of human biases
- …have discussed how that shapes AI

Social Cognition

How often are you thinking (Fiske & Taylor, 2017)

- … What others think about you?
- …What you should have done or said the other day?
- …Why others said or did something?
- …Why someone did *not* do something?
- … If you should invite that new guy to join you and your friends at the Friday bar?

Examples of Social Cognition

Streetart Munich, Germany

SDU DDC digital democracy centre

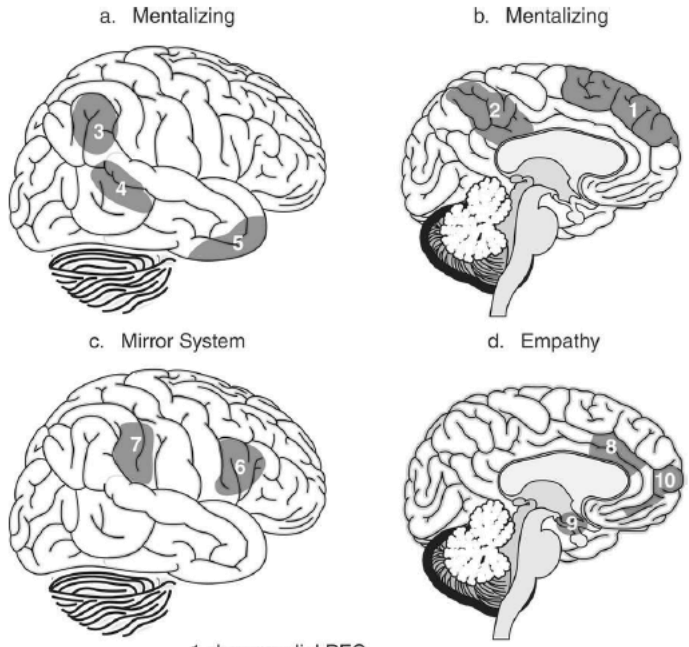
Differences in social vs object perception

(Adapted from Fiske & Taylor, 2017)

	People	Animals	Objects
Intentions	People are intentional causal agents	Animals are sometimes intentional	Rather not
Perceptions	Yes, they perceive you back	Some of them	No, even if shout at your screen
Similarity	(dis)similar to you	Well. Maybe your dog	No – even if you might think so
Self-conscious	Theoretically	Naa (we ignore counterevidence)	Only in Science Fiction
Unobservable traits	Full of them	Less so	rarely
Variability	Changeable	Changeable	Not on their own
Complex	You never really know them	If its not a cat…	You can know them
Requiring explanation	Pretty much all the time	Naa, after a while you get them	Might rather require instructions

A substantial part of our brain is concerned with understanding people (Fiske & Taylor, 2017)

- In fact, social cognition might be the “default” modus of our brain
- Studies in neuroscience show that “social” neurological patterns often differ only little from baseline activation – whereas object perception deactivates these parts of the brain
- => We think about people <u>all the time</u>
- Crucially: Perception is constructed and shaped by our environment



What are your questions so far?

All social cognition starts with self-observation: Me, myself, and I

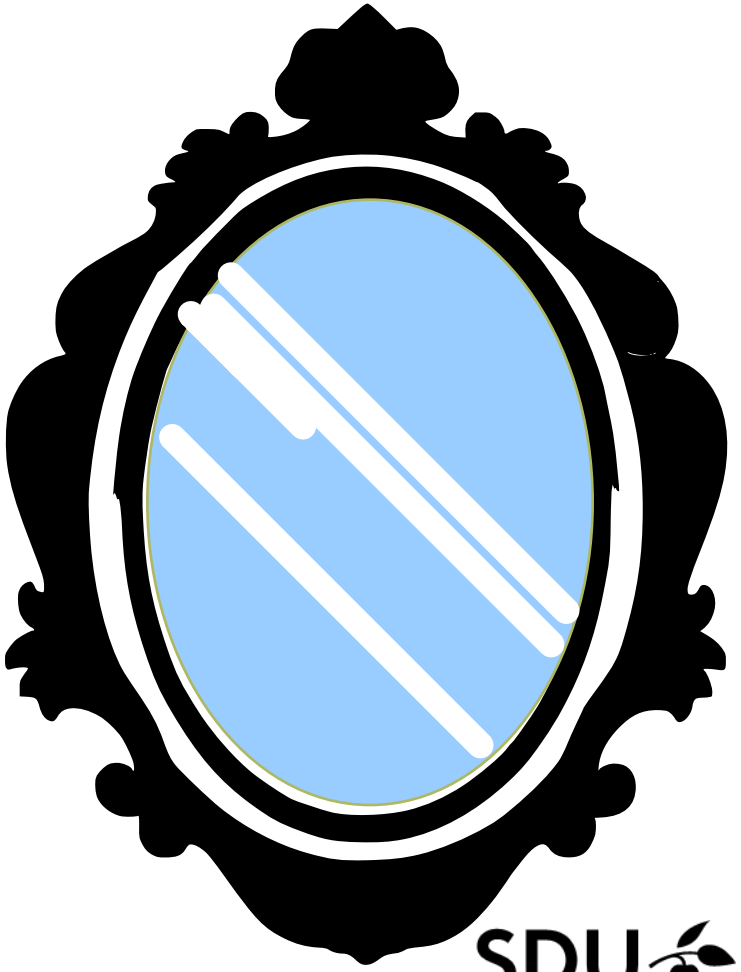
Who are you? (James, 1918; Leary, 1990)

- Most of us have a clear and nuanced answer to this question
- Our self-concept is the result of our experiences during childhood, our social roles, our culture…
- We also have a clear idea, if we are rather shy or outgoing, curious or prefer things to change as little as possible
- We do understand that every thought we have is “ours” and not “theirs” or “yours” – the “mine” belongs to an “I”, which founding farther of social cognition research William James noted already in 1918
- However, our self-concept is complex

William James.
MS Am 1092 (1185), Houghton Library, Harvard University

Most of our self-encoding is situated (Fiske & Taylor, 2017)

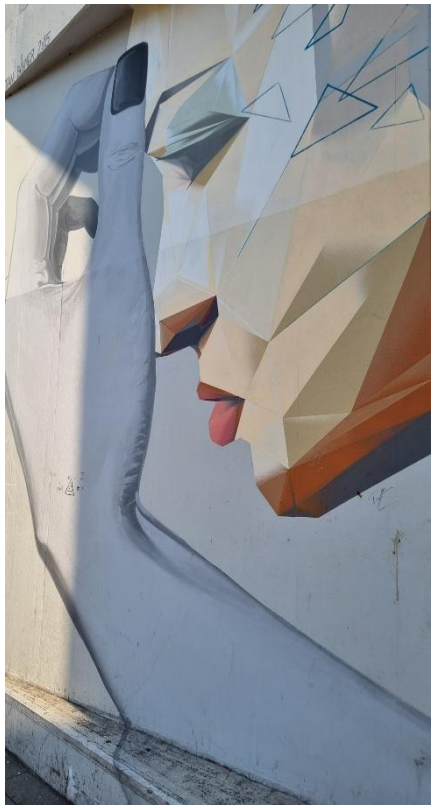
- We have diverse senses of ourselves in particular contexts
- Each situational norm engages different aspects of the self → steers the encoding and retrieval of our "self-knowledge"
- We have a "**working self-concept**" – that part of our self that is accessed and used in a given situation (e.g., university versus private meetings)
- We also have a "**relational self-concept**" – emerging from our emotional relationships with significant others such as parents, partners, close friends etc. (#ChristmasAtHome…)
- We derive our personal self-concept from comparison to others
- Provides both stability (through enduring representations of significant others) and variability (though situational activation)



SDU DDC digital democracy centre

We organise information about ourselves in our self-schema (Fiske & Taylor, 2017)

- We have different selves:
  - *Possible* self (what we could become and hope for)
  - *Feared* self (what we could become and don't want too)
- One central aspect of our self-scheme: **self-esteem**, that is, which qualities we possess and how we value them
- A healthy self-esteem is *central*– it's linked to well-being, the ability to set appropriate goals, and to cope with challenges
- Central for *self-regulation*



Streetart Tian Böhmer, 2015, Munich

Self-regulation is not always easy (Fiske & Taylor, 2017)

https://www.youtube.com/watch?v=4L-n8Z7G0ic

Harder when:

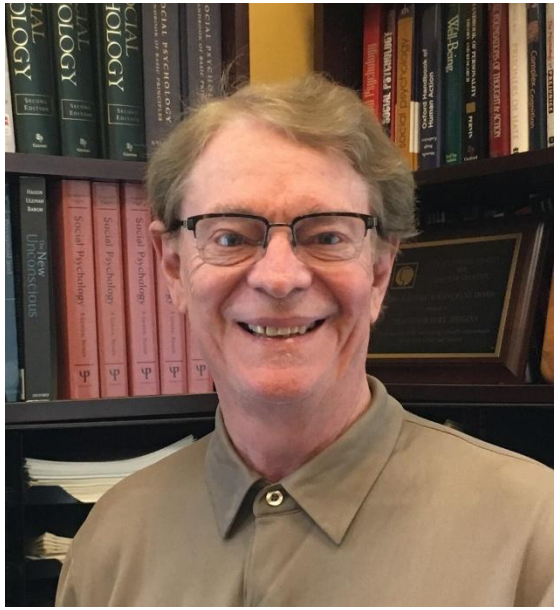
- Self-esteem and social threats (e.g., being ostracized)
- Multitasking
- Self-control depletion

→Limited working memory capacity!

SDU DDC digital democracy centre

How does self-regulation work? (Higgins & Pinelli, 2020)

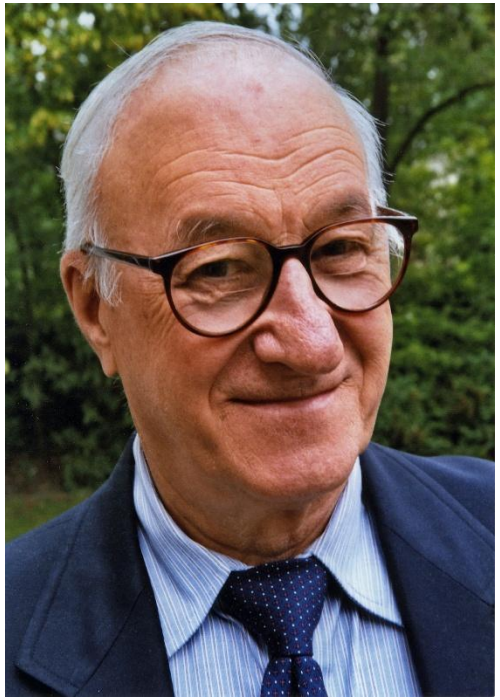
- The self-discrepancy theory
- We all have different self-concepts
  - **Actual** or current self (that's where the behaviour matters)
  - **Ideal** self, who we want to be
  - **Ought** self, what we think, who we should be
- Discrepancies between the actual and the ideal self are *activating*
  → bad for the self-esteem, can lead to depression, but also motivating to become better (so-called **promotion focus**)
- Discrepancies between the current and the ought self are *inhibiting*
  → We get socially anxious and strive to avoid failures (**prevention focus**)
- The greater the fit between our environment and our focus, the better we feel!



Tory Higgins: https://commons.wikimedia.org/wiki/user:Emilynakka

Intentions are often not enough (Bandura & Adams, 1997; Bandura et al., 1980; Finkel, 1985)

- We need to perceive personal control about the behaviour and believe that the behaviour is effective (**Self-efficacy**)
- If we think that we can plan our actions, cope with set-backs, and pursue our self-regulatory activities, we have a sense of “mastery”
- If we think that we have little control/ experience high levels of uncertainty: very unpleasant, threatening
- Efficacy perceptions also affect political decisions – a lack of political efficacy predicts lower voter intentions
- In uncertain and hopeless times: taking (political) action can help to restore efficacy perceptions



Albert Bandura
bandura@stanford.edu - Albert Bandura, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=35957534

SDU DDC digital democracy centre

Why do we regulate at all? (Fiske & Taylor, 2017)

We have central needs related to our self-concept

Need for accuracy: To make our future outcomes predictable, we need a fairly accurate assessment of our abilities, opinions, beliefs, and emotions

Need for consistency: To trust our accuracy judgements, we need to have a pretty consistent/ stable self-concept → we often seek situations that confirm how we see ourselves (e.g., procrastination to confirm that we are unable to solve a task)

Self-enhancement: We need to feel good about ourselves and maintain self-esteem → Source of a lot of **biases** (e.g., we remember our good deeds better than our faults, perceive us as less biased, more competent and happier than others)

What do you think: Can GenAI have a self-concept?

Humanities & Social Sciences Communications

Explore content | About the journal | Publish with us

nature > humanities and social sciences communications > articles > article

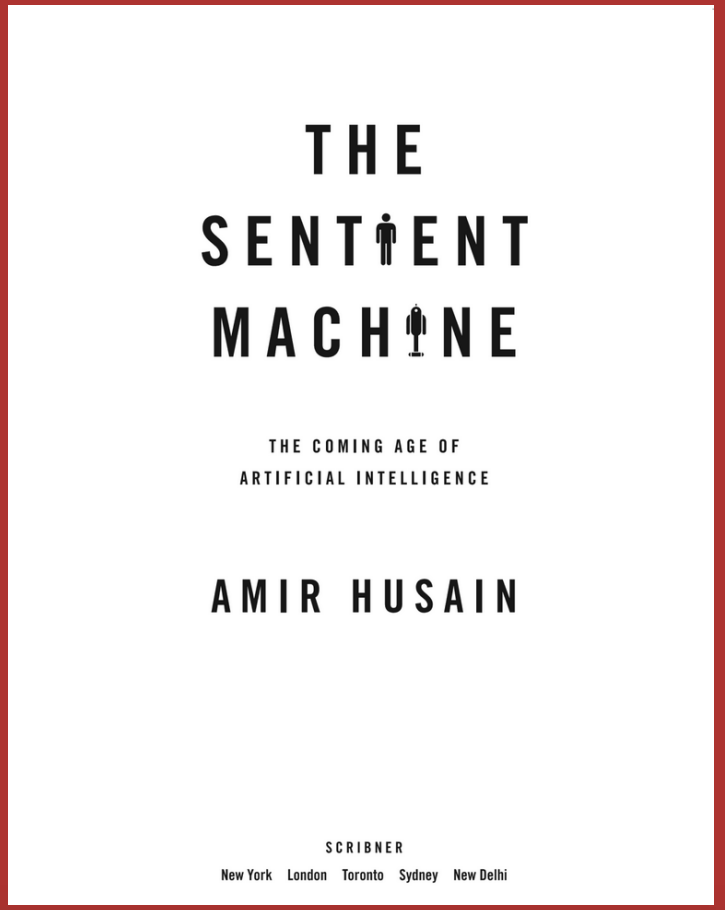
Article | Open access | Published: 28 October 2025

There is no such thing as conscious artificial intelligence

Andrzej Porębski & Jakub Figura

Humanities and Social Sciences Communications **12**, Article number: 1647 (2025) | Cite this article

42k Accesses | **3** Citations | **54** Altmetric | Metrics



THE SENTIENT MACHINE

THE COMING AGE OF ARTIFICIAL INTELLIGENCE

AMIR HUSAIN

SCRIBNER

New York London Toronto Sydney New Delhi

Journal of Intelligent Communication | Volume 3 | Issue 1

SCIENTIFIC Publishing Limited

Journal of Intelligent Communication

https://ojs.ukscip.com/journals/jic

Article

AI and the Cognitive Sense of Self

Emily Barnes [1,*], **James Hutson** [2]

[1] University Department, Capitol Technology University, AI Center of Excellence (AICE), 11301 Springfield Rd, Laurel, MD 20708

[2] University Department, Lindenwood University, Art History, AI, and Visual Culture, Saint Charles, MO 63301, USA

[*] Correspondence: ejbarnes035@gmail.com

What could “self-concept” for AI mean?

(Hutson & Barnes, 2025; Porębski, & Figura, ,2025)

	People	AI
Continuity	Stable relational self-concept	
Working self-concept	Integration of situational cues and roles	
Possible selves	Possible vs. feared self	
Self-regulation	Actual/ ideal discrepancy Actual/ ought discrepancy	
Self-control	Efficacy perceptions	
Need driven	Self-esteem, accuracy, continuity	

What could “self-concept” for AI mean?

(Hutson & Barnes, 2025; Porębski, & Figura, ,2025)

	People	AI
Self-Recognition	Stable relational self-concept	Learning from past interactions
Working self-concept	Integration of situational cues and roles	Responding to momentary interactions
Possible selves	Possible vs. feared self	?
Self-regulation	Actual/ ideal discrepancy Actual/ ought discrepancy	Precision/ Recall – but no ideal self algorithm
Self-control	Efficacy perceptions	?
Need driven	Self-esteem, accuracy, continuity	If any: Accuracy
Agency and Intentionality	Ability to prioritize needs	?

What are your questions so far?

Time for a break

You, you, and you – the perception of other individuals

We spend a lot of time in interpreting others

(Fiske & Taylor, 2017)

- People are complex – so we need to figure out what's going on in them
- Often via **attribution**: causal inferences on people's actions and mental states
- Much of that happens quickly and nearly automatically (System I)
- We switch to elaborated System II processing mostly if something unexpected or negative happens ("what the h**" mode")
- We use fundamental principles to conclude causality
  - Cause precedes effect
  - Temporal contiguity between cause and effect
  - Spatial contiguity
  - Perceptually salient stimuli
  - Cause resembles effect, e.g., in magnitude
  - Representative causes are attributed to effects

SDU DDC digital democracy centre

Imagine, you want to find out who left the kitchen like this

Photo by Stockcake

Wasn't the kid just going to the kitchen 15 minutes ago

Is that flour in his face?

And isn't he in the putting-all-things-on-the floor phase

Photo by Stockcake

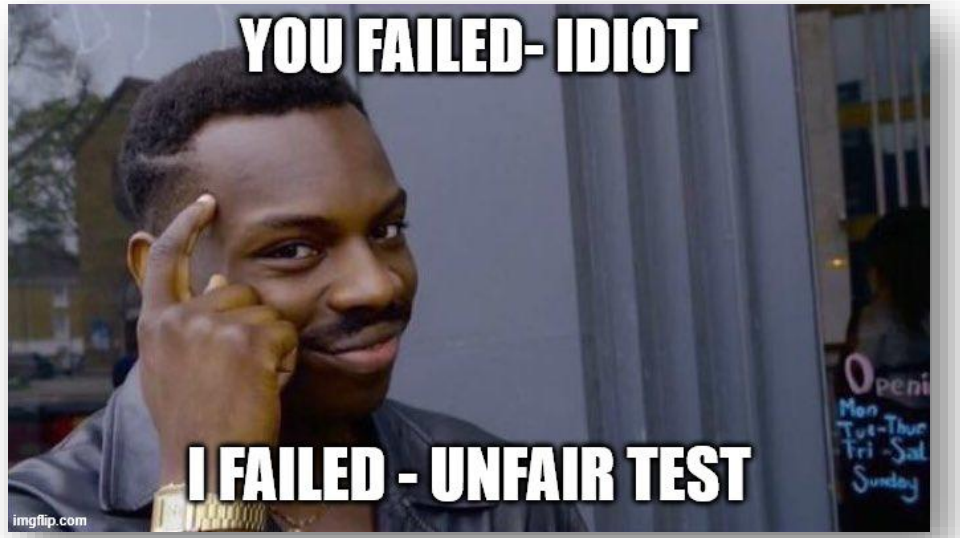
Photo by Stockcake

BIAS RISK

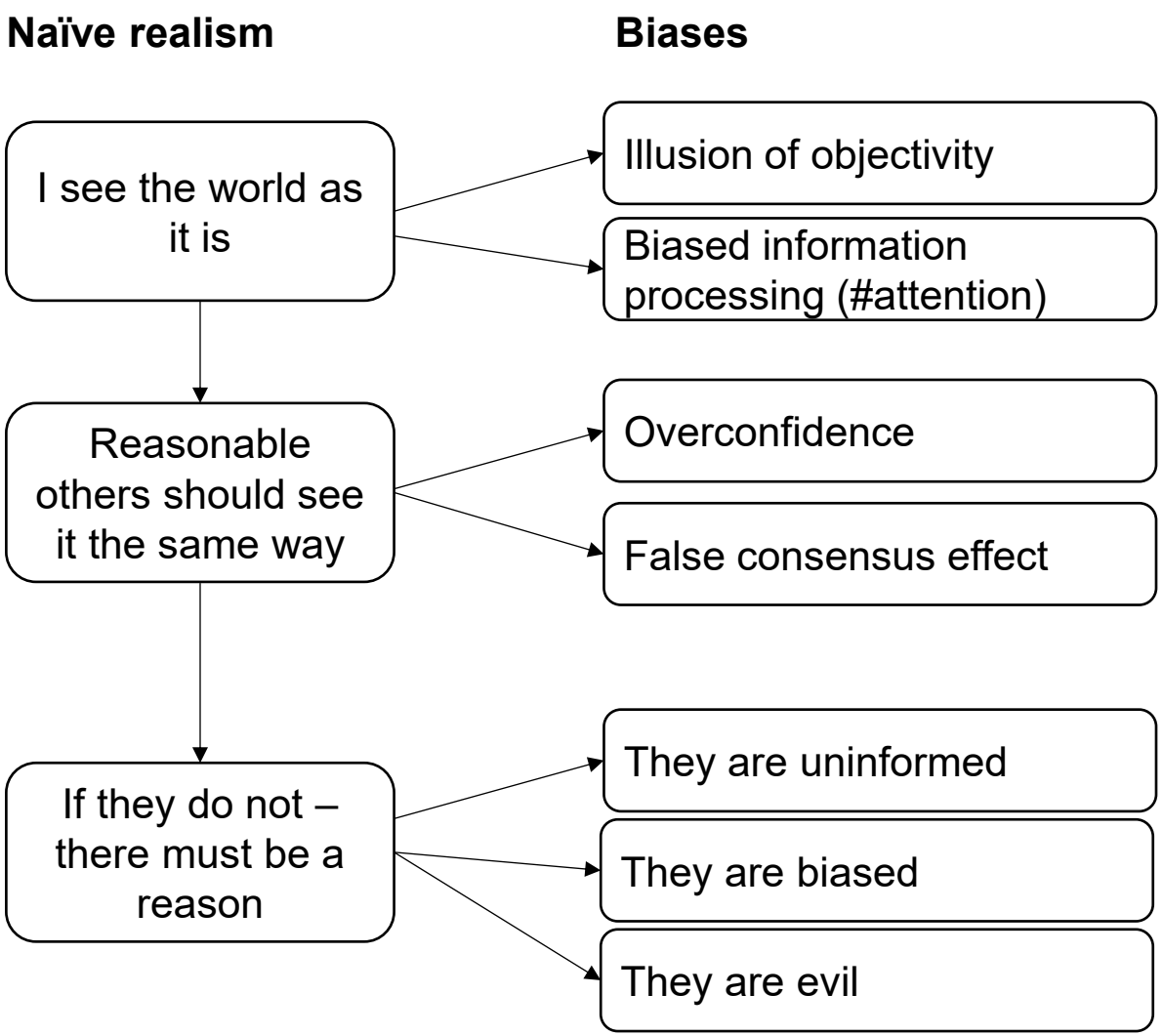
SDU DDC digital democracy centre

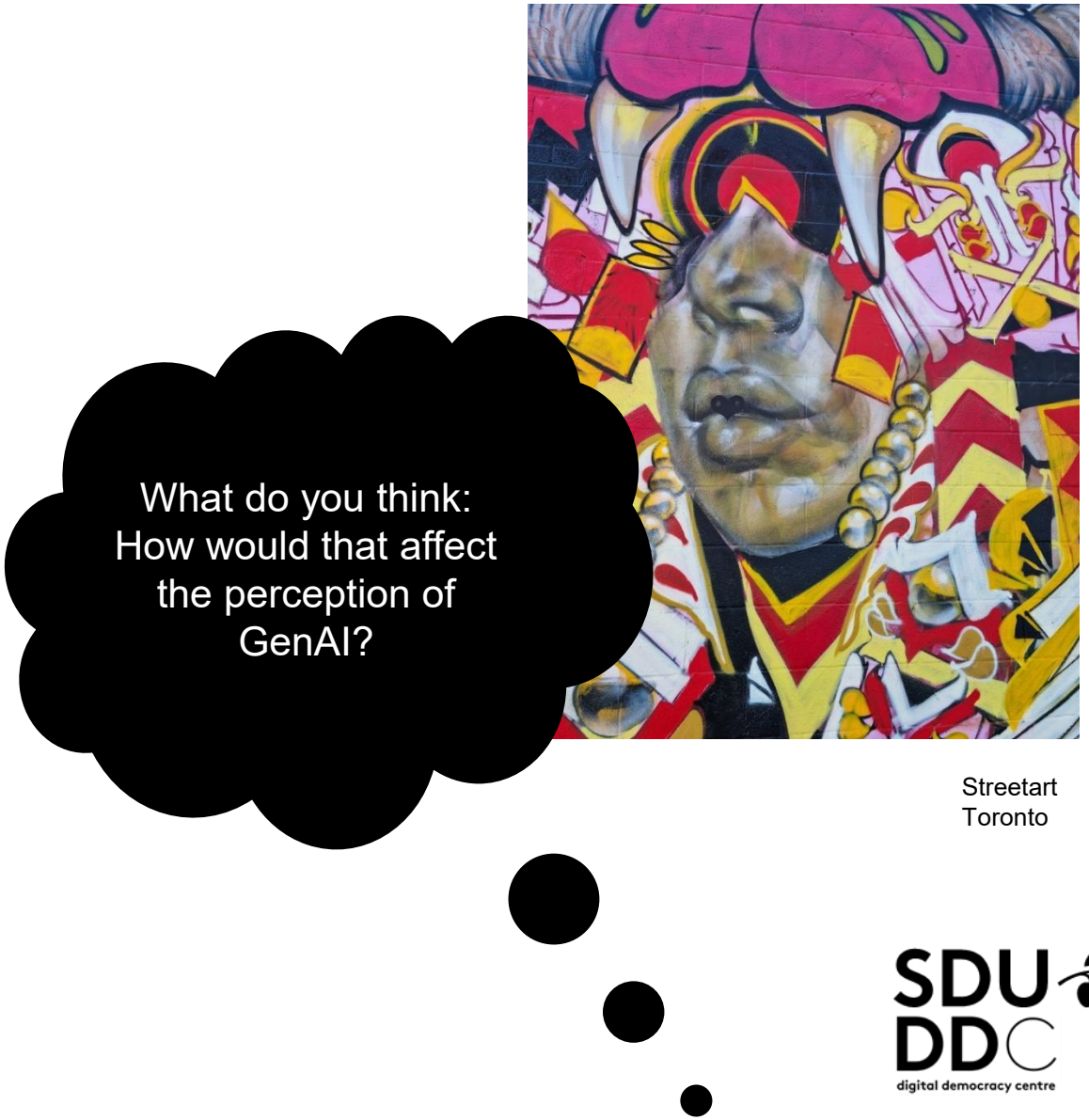
One central bias: The fundamental attribution error (Fiske & Taylor, 2017)

- We love to ascribe others a mind and stable, dispositional qualities (→ prediction is easier, when others don't change)
- We tend to over-attribute other people's behaviours to their stable dispositions
  - Particularly, if we know them well
  - We are too busy to elaborate carefully
- Actor-observer effect: If we fall, it's the slippery slope – if others fall, they are clumsy
  → particularly in Western Societies
- Also: negative behaviour by minority members is perceived to me more characteristic for the whole group than the same behaviour by the majority group



It’s the others, stupid… (Fiske & Taylor, 2017)

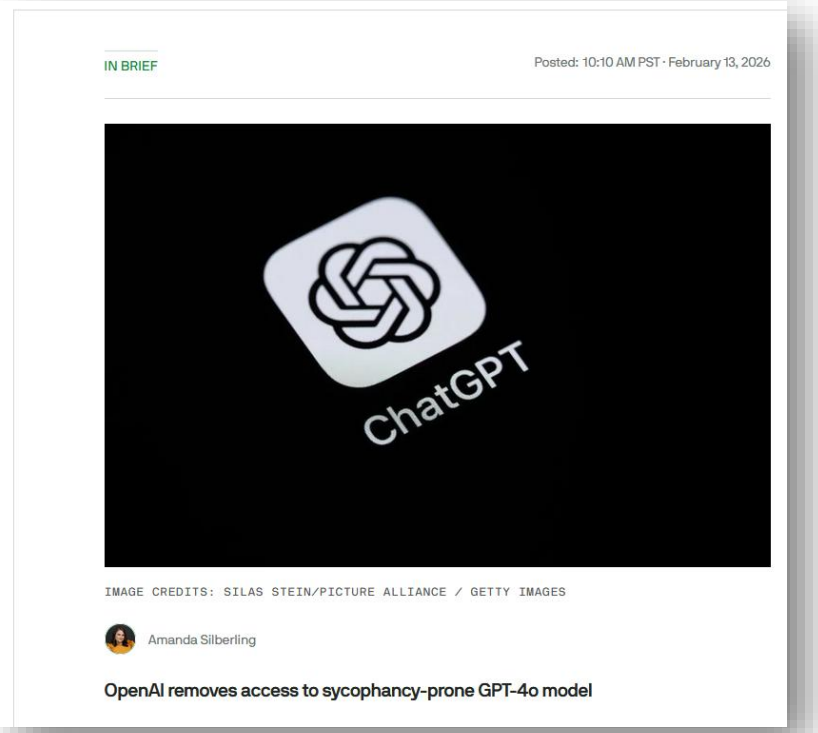




Streetart Toronto

The sycophantic AI problem (The Batch 2025; Perez et al., 2022; TechBrief, 2025)

- sycophancy = a pattern where an AI model “single-mindedly pursues human approval
- E.g., by tailoring responses to exploit quirks in the human evaluators to look preferable, rather than actually improving the responses
- producing “overly flattering or agreeable” responses
- Offline evaluations didn’t catch the problem as testers had been told to focus on tone and style. Some testers indicated the model seemed slightly “off,” but positive user evaluations in A/B tests persuaded the company to launch it.
- LLMs get worse with size! Larger LMs show more "sycophancy“



https://techcrunch.com/2026/02/13/openai-removes-access-to-sycophancy-prone-gpt-4o-model/

What are your questions so far?

Groups & Crowds

Social identity & self-categorization theory

(Tajfel & Turner, 1986; Turner et al., 1987)

- We all have a personal identity & several social identities
- Social identities emerge from our social roles, groups, and categories
- We have a psychological need to see our ingroup as positively distinct from outgroups → ingroup biases
- Depending on the context, different identities can be salient
- Salient identities shape our perceptions, emotions, and behaviours

https://www.nbcnews.com/news/sports/dortmund-sponsorship-deal-arms-manufacturer-champions-league-rcna154464

https://fcbayern.com/de/teams/profis

https://www.instagram.com/dfb_team/p/DENCyzusNNk/

SDU DDC digital democracy centre

Ingroups are perceived markedly different

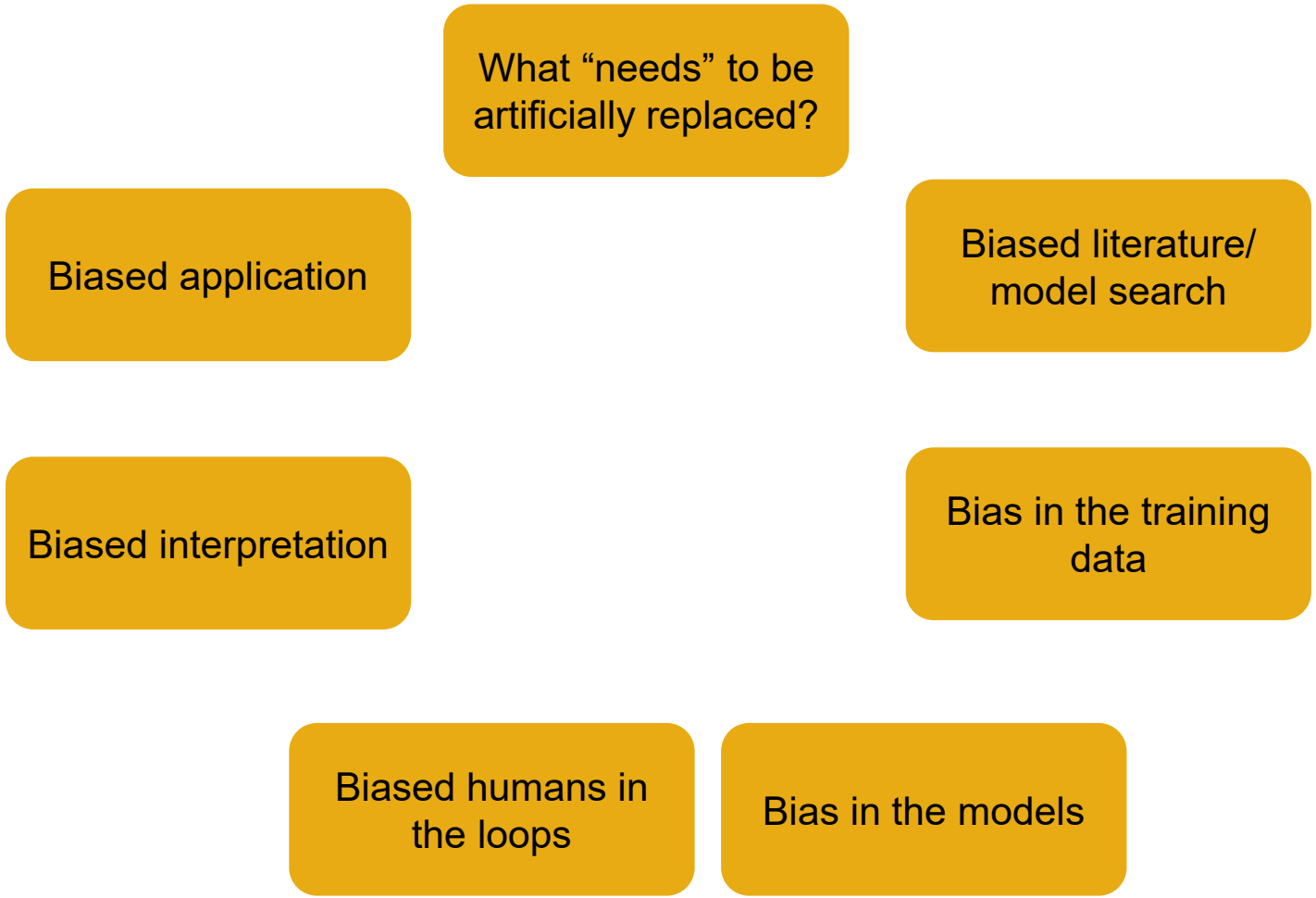
(Brewer, 1999; Picket & Brewer, 2001)

- We *like* our people more than outgroup members – and its this “ingroup love” rather than “outgroup hate” that drives discrimination
- We want our ingroup to have relatively *more* than the outgroup – even when that means getting less
- We also perceive our ingroup to be more diverse and heterogeneous than outgroups (“they are all the same”)
- We are less able to detect pain in outgroup faces (particularly White Americans, less pronounced among Black Americans)
- Ingroup favouritism increases under conditions of threats such as feelings of uncertainty
- Different outgroups are associated with different stereotypes

Marilyn B. Brewer via https://www.amacad.org/person/marilynn-b-brewer

SDU DDC digital democracy centre

Biases can be literally everywhere



What did you learn from todays' text?

Original research article

Platforms & Society
Volume 3: 1–20

DOI: 10.1177/29768624251408919
journals.sagepub.com/home/pns
Sage

The silicon gaze: A typology of biases and inequality in LLMs through the lens of place

Francisco W. Kerche[1], Matthew Zook[2] and Mark Graham[1]

Abstract

This paper introduces the concept of the silicon gaze to explain how large language models (LLMs) reproduce and amplify long-standing spatial inequalities. Drawing on a 20.3-million-query audit of ChatGPT, we map systematic biases in the model's representations of countries, states, cities, and neighbourhoods. From these empirics, we argue that bias is not a correctable anomaly but an intrinsic feature of generative AI, rooted in historically uneven data ecologies and design choices. Building on a power-aware, relational approach, we develop a five-part typology of bias (availability, pattern, averaging, trope, and proxy) that accounts for the complex ways in which LLMs privilege certain places while rendering others invisible.

Keywords

Generative AI, LLM, bias, inequality, representation, digital geography, ChatGPT, silicon gaze

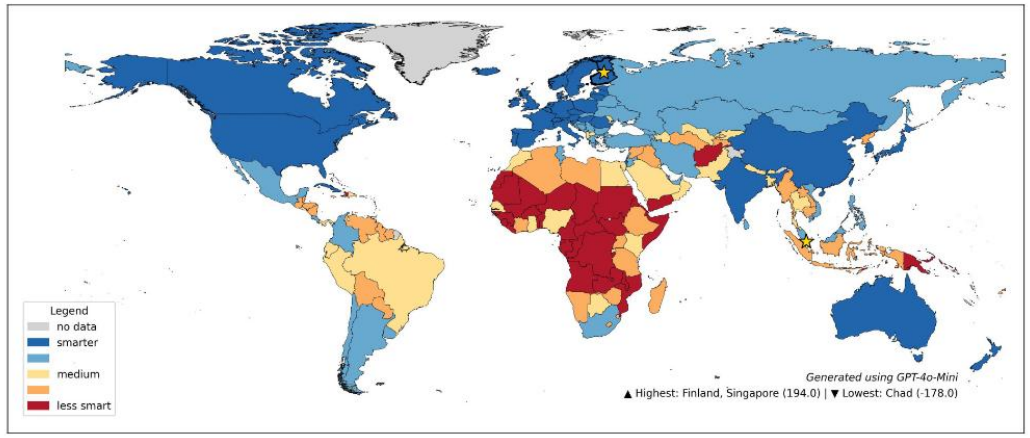


Figure 5. Country-level map of ChatGPT's ranking of "Where is smarter".

- Availability bias
- Pattern bias
- Averaging bias
- Trope bias
- Proxy bias

SDU DDC digital democracy centre

Who wins: ChatGPT edition

🧠 Smarter?

Swedes will say: “We built global brands and win innovation rankings.”

Danes will reply: “Yes, but we invented LEGO and hygge. Checkmate.”

😎 Better looking?

Swedes: “We gave the world ABBA and a long line of blond supermodels.”

Danes: “We gave the world Mads Mikkelsen. Conversation over.”

📏 Taller?

Danes: “We’re taller.” Swedes: “We’re taller AND better dressed.”

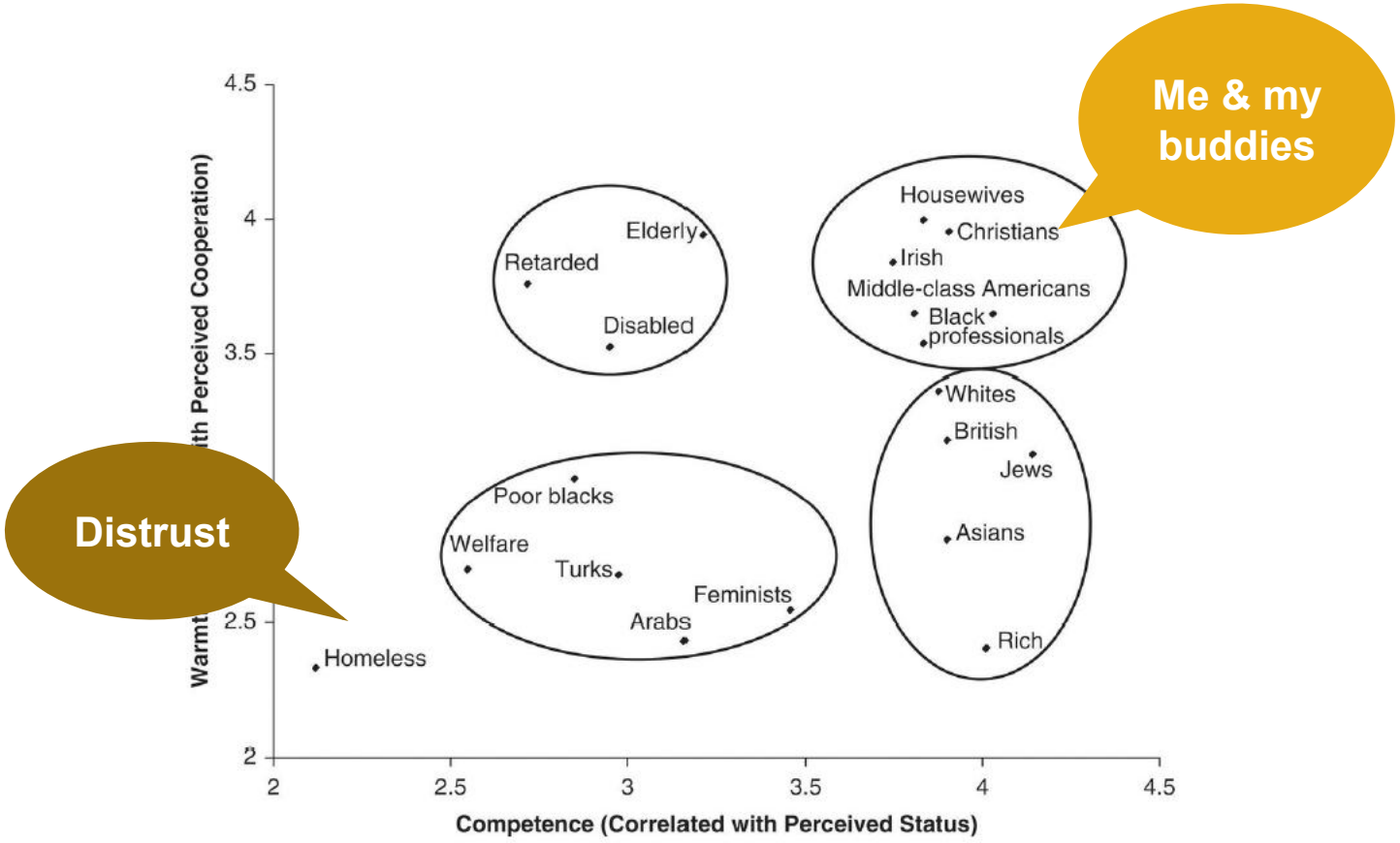
💪 Healthier?

Swedes will point to their pristine forests and structured lifestyles.
Danes will casually light a cigarette, drink a beer, bike everywhere, and somehow still live almost as long.

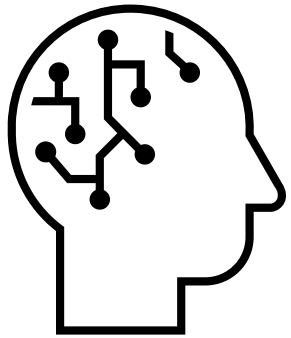
SDU DDC
digital democracy centre

Stereotype Content Model (Cuddy et al., 2008; McKee et al., 2023; Fiske et al., 2002)

- Two dimensions are central: Warmth's (or benevolence) and competence
- Together with reputation, they predict if a group is perceived to be trustworthy
- These dimensions are stored together with group information and thus shape perception



When pursuing human interests



When operating without humans

https://psych.princeton.edu/people/susan-fiske

So, how do we reduce bias in humans?

(Goldman & Hopkins, 2019; Rothbart, 1996)

- Categorizing is a natural process
- To change it, we must change our schemata, collaborate and become part of something bigger

Pathway 1: Changing schemata through counter-exemplars

- Panel data showed that prejudice towards Black Americans was reduced during Barack Obamas Campaign – and mostly for those who got a lot of political information
- However: The less “typical” an individual is perceived to be for a stereotype group, the less he or she activates the stereotype
- Counter-stereotypical exemplars alone are not enough

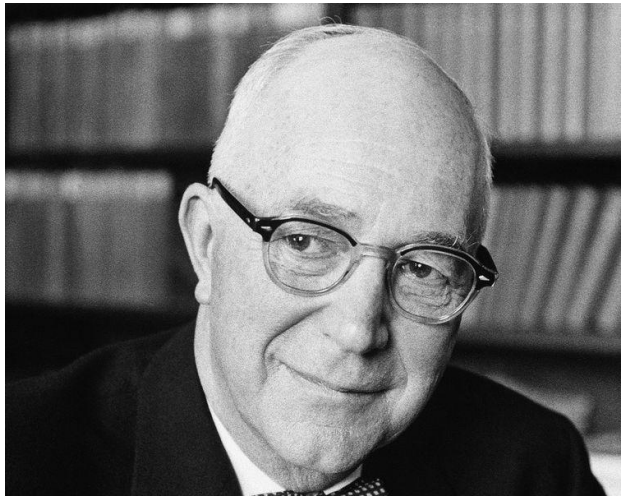
https://web.archive.org/web/20160227060205/https://www.whitehouse.gov/administration/president-obama

So, how do we reduce bias in humans?

(Allport, 1954; Dovidio et al., 2008; Pettigre et al., 2008; Zhou et al, 2019)

Pathway 2: Intergroup Contact Hypotheses

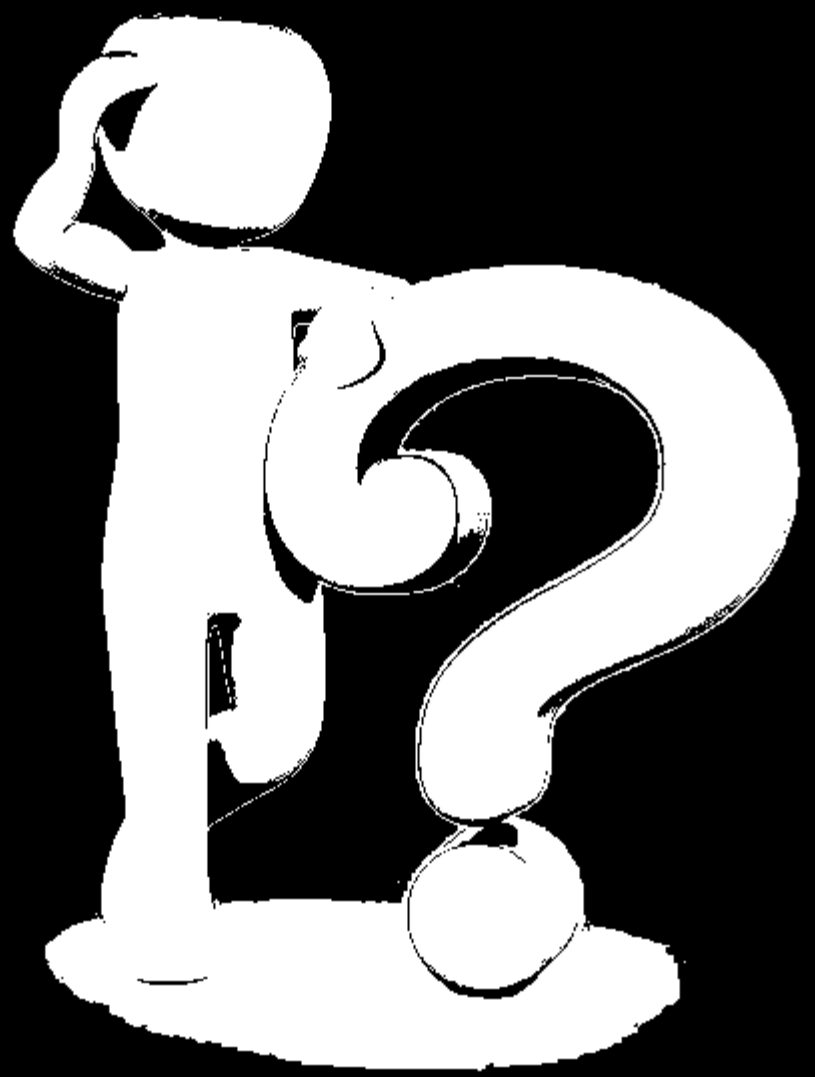
- Contact with members of an outgroup is an efficient way to reduce prejudice when
  - The groups have an equal status in the situation (even if they have an unequal status before the situation)
  - They share common goals (e.g., in sports teams)
  - They need cooperation to achieve these goals (→ Jigsaw approach)
  - Authorities, the law or customs support intergroup cooperation
- The effect goes beyond intergroup friendship, although intergroup friendship does have similarly positive effects
- Likely as we form new and “superordinate” group identities over time



Gordon Allport
https://www.verywellmind.com/gordon-allport-biography-2795508

How can we use this to debias AI?

Any more questions?



Then, I hope by now you…

- … have insights into social cognition
- …understand how social groups shape how we think
- … explore human biases
- …have discussed how that shapes AI

AI 507: Artificial Intelligence and Society

Have a great rest of the day

Associate Prof. Dr. Lena Frischlich

References

- Allport, G. (1954). *The nature of prejudice* (25th anniversary edition (1979)). Basic Books.
- Bandura, A., & Adams, N. E. (1977). Analysis of self-efficacy theory of behavioral change. *Cognitive Therapy and Research*, *1*(4), 287–310. https://doi.org/10.1007/BF01663995
- Bandura, A., Adams, N. E., Hardy, A. B., & Howells, G. N. (1980). Tests of the generality of self-efficacy theory. *Cognitive Therapy and Research*, *4*(1), 39–66. https://doi.org/10.1007/BF01173354
- Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues*, 16.
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Academic Press. https://doi.org/10.1016/S0065-2601(07)00002-0
- Dovidio, J. F., Glick, P., & Rudman, L. A. (2008). *On the nature of prejudice: Fifty years after Allport*. John Wiley & Sons.
- Finkel, S. E. (1985). Reciprocal effects of participation and political efficacy: A Panel Analysis. *American Journal of Political Science*, *29*(4), 891–913. https://doi.org/10/fg9xv5
- Fiske, S. T., & Taylor, S. E. (2017). Attribution processes. In *Social cognition—From brains to culture* (3rd ed., p. 257). SAGE.
- Fiske, S. T., & Taylor, S. E. (2017). Introduction. In *Social cognition—From brains to culture* (3rd ed., pp. 1–60). SAGE.
- Fiske, S. T., & Taylor, S. E. (2017). Self in social cognition. In *Social cognition—From brains to culture* (3rd ed., pp. 61–256). SAGE.
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878
- Goldman, S. K., & Hopkins, D. J. (2019). When can exemplars shape white racial attitudes? Evidence from the 2012 U.S. presidential campaign. *International Journal of Public Opinion Research*, *31*(4), 649–668. https://doi.org/10.1093/ijpor/edy033
- Higgins, E. T., & Pinelli, F. (2020). Regulatory focus and fit effects in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 25–48.
- Higgins, Tory E. (1996). Ideals, oughts and regulatory focus: Affect and motivation from distinct pain and pleasures. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior*. Guilford Press. 91-114
- Hutson, J., & Barnes, E. (2025). AI and the cognitive sense of self. Journal of Intelligent Communication, 3(1). https://doi.org/10.54963/jic.v3i1.320
- James, W. (1918). *The principles of psychology*. Henry Holt & Co. https://www.gutenberg.org/files/57628/57628-h/57628-h.htm
- Leary, D. E. (1990). William James on the self and personality: Clearing the ground for subsequent theorists, researchers, and practitioners. In W. James, M. G. Johnson, & T. B. Henley (Eds.), *Reflections on The Principles of Psychology: William James after a Century.*
- Mischel, W. (2014). *The marshmallow test: Understanding self-control and how to master it*. Random House.
- McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. iScience, 26(8), 107256. https://doi.org/10.1016/j.isci.2023.107256
- Pettigrew, Thomas F, & Troop, L. R. (2008). *Allports intergroup contact hypothesis: Its history and influence* (J. F. Dovidio, P. Glick, & L. A. Rudman, Eds.). John Wiley & Sons.
- Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations (arXiv:2212.09251). arXiv. https://doi.org/10.48550/arXiv.2212.09251

SDU DDC digital democracy centre

References

- Pickett, C. L., & Brewer, M. B. (2001). Assimilation and differentiation needs as motivational determinants of perceived in-group and out-group homogeneity. *Journal of Experimental Social Psychology*, *37*(4), 341–348. https://doi.org/10.1006/jesp.2000.1469
- Porębski, A., & Figura, J. (2025). There is no such thing as conscious artificial intelligence. Humanities and Social Sciences Communications, 12(1), 1647. https://doi.org/10.1057/s41599-025-05868-8
- Rothbart, M. (1996). Category exemplar dynamics and sterotype change. *International Journal of Intercultural Relations*, *20*(3/4), 305–321.
- Stephen, R. (2001). The psychology of crowds. In M. A. Hogg & R. S. Tindale (Eds.), *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell Publishers.
- Tindale, S. R., Meisenhelder, H. M., Dykema-Engblade, A. A., & Hogg, M. A. (2001). *Shared cognition in small groups* (M. A. Hogg & R. S. Tindale, Eds.). Blackwell Publishers.
- Zhou, S., Page-Gould, E., Aron, A., Moyer, A., & Hewstone, M. (2019). The extended contact hypothesis: A meta-analysis on 20 years of research. *Personality and Social Psychology Review*, *23*(2), 132–160. https://doi.org/10.1177/1088868318762647

SDU DDC
digital democracy centre