Unpacking the Ethical Implications of Bias and Data Use in Generative AI Models

conversation with Safiya Noble & Cecil Rosner

Rod Lastra
University of Manitoba

#5

Safiya U. Noble Professor of University of California, Los Angeles

Cecil Rosner is a journalist, writer and adjunct professor

OBJECTIVE TRUTHS

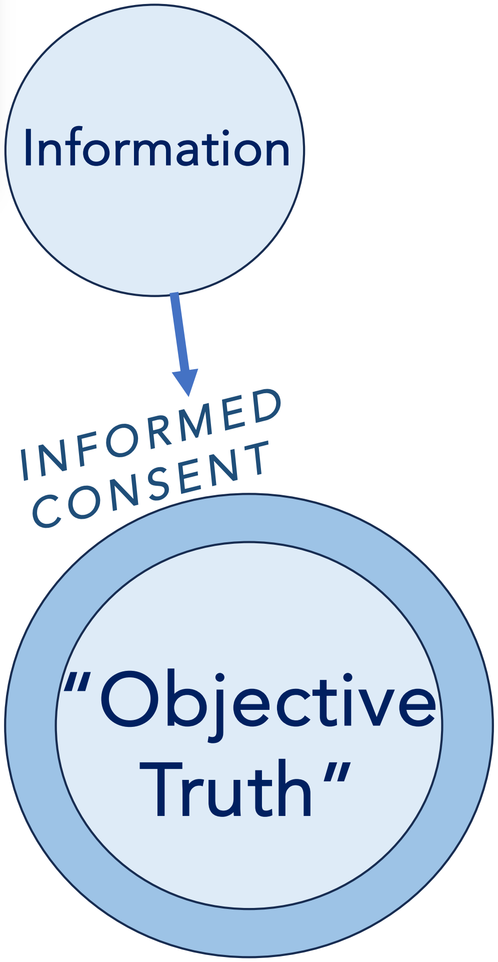
and the

“ENGINEERING OF CONSENT“

“ENGINEERING OF CONSENT”

“Objective Truth”

“ENGINEERING OF CONSENT”

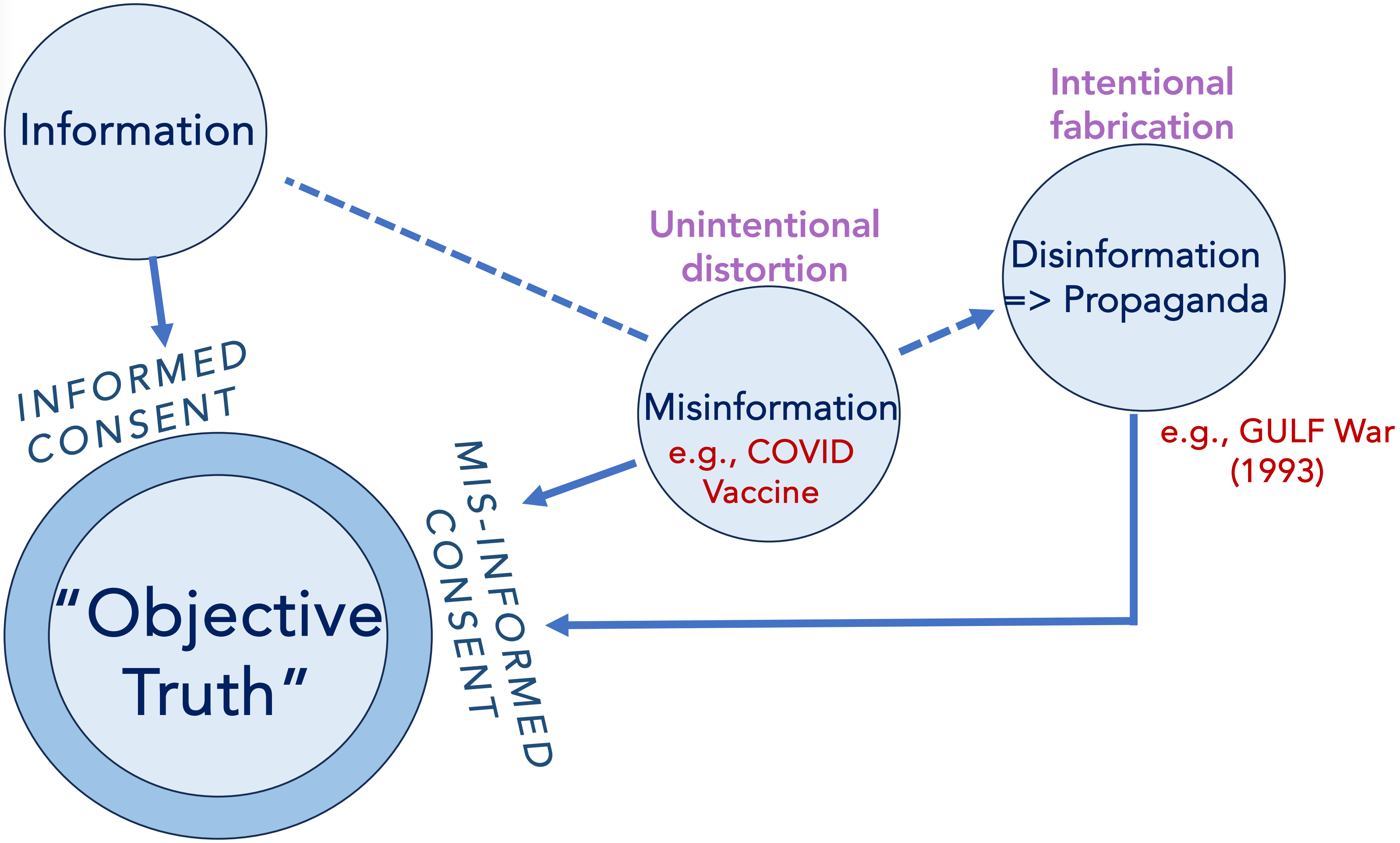


Information

INFORMED CONSENT

“Objective Truth”

“ENGINEERING OF CONSENT”



Information

INFORMED CONSENT

“Objective Truth”

Unintentional distortion

Misinformation

e.g., COVID Vaccine

MIS-INFORMED CONSENT

Intentional fabrication

Disinformation => Propaganda

e.g., GULF War (1993)

“HUMAN PERCEPTIONS OF TRUTH“ Science of manipulation

Key publications (examples):

- Arkes, H. R., Boehm, L. E., & Xu, G. (1991). **Determinants of judged validity**. Journal of Experimental Social Psychology, 27(6), 576–605. https://doi.org/10.1016/0022- 1031(91)90026-3

- Bacon, F. T. (1979). **Credibility of repeated statements**: Memory for trivia. Journal of Experimental Psychology: Human Learning & Memory, 5(3), 241–252.

- Beck, J. (2017). This article won’t change your mind: **The fact on why facts alone can’t fight false beliefs**. The Atlantic. Retrieved from https://www.theatlantic.com/science/archive/2017/03/this-article-wont-change-your- mind/519093/

- Hasher, L. D. Goldstein, and T. Toppino (1977). **Frequency and the conference of referential validity**. Journal of verbal behavior, 16, 107-112

- Schwartz, M. (1982). **Repetition and Rated Truth Value of Statements**. The American Journal of Psychology, 95(3), 393–407. https://doi.org/10.2307/1422132

- Van der Linden, S. (2023). Foolproof: **Why misinformation infects our minds and how to build immunity**. WW Norton & Company.

- Vosoughi, S., Roy, D., & Aral, S. (2018). **The spread of true and false news online**. *science*, *359*(6380), 1146-1151.

“HUMAN PERCEPTIONS OF TRUTH“

Science of manipulation

“Human mind is suspectable to influence “

“falsehood and truths blend and persists, influencing how we perceive reality“

Key publications (examples):

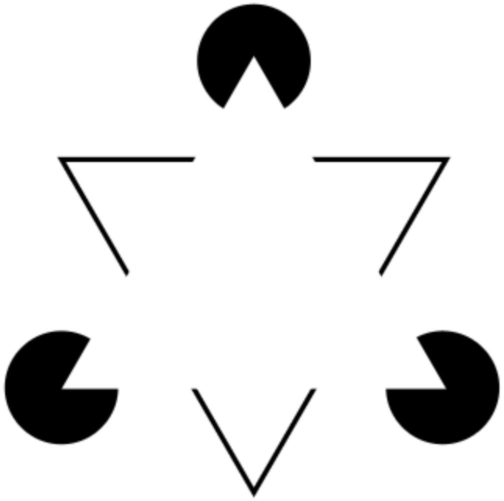
- Ar[illegible]. Journal of Experimental Sc[illegible]6), 576–605. https://doi.org/10.1016/0022-1031(91)90026-3
- Bacon, F. T. (1979). **Credibility of repeated statements**: Memory for trivia. Journal of Experimental Psychology: Human Learning & Memory, 5(3), 241–252.
- Beck, [illegible] **can't fight false beliefs**. The Atlantic. Retrieved from https://www.theatlantic.com/science/archive/2017/03/this-article-wo[illegible]d/5[illegible]
- Hasher, L., D. Goldstein, and T. Toppino (1977). **Frequency and the conference of referential validity**. Jo[illegible]ehavior, 16, 107-112
- Schwartz, M. (1982). **Repetition and Rated Truth Value of Statements**. The American Journal of Psychology, 95(3), 393–407. https://doi.org/10.2307/1422132
- Van der Linden, S. (2023). Foolproof: **Why misinformation infects our minds and how to build immunity**. WW Norton & Company.

“HUMAN PERCEPTIONS OF TRUTH”
Science of manipulation

PERCEPTION OF TRUTH

- **Informed by previous experiences**
- Our minds fill our perception
- “**ILLUSIONARY TRUTH EFFECT**”

- **REPETITION** – more familiar the claim, easier to process
  - Fake news: simple
  - Science: complex and nuanced



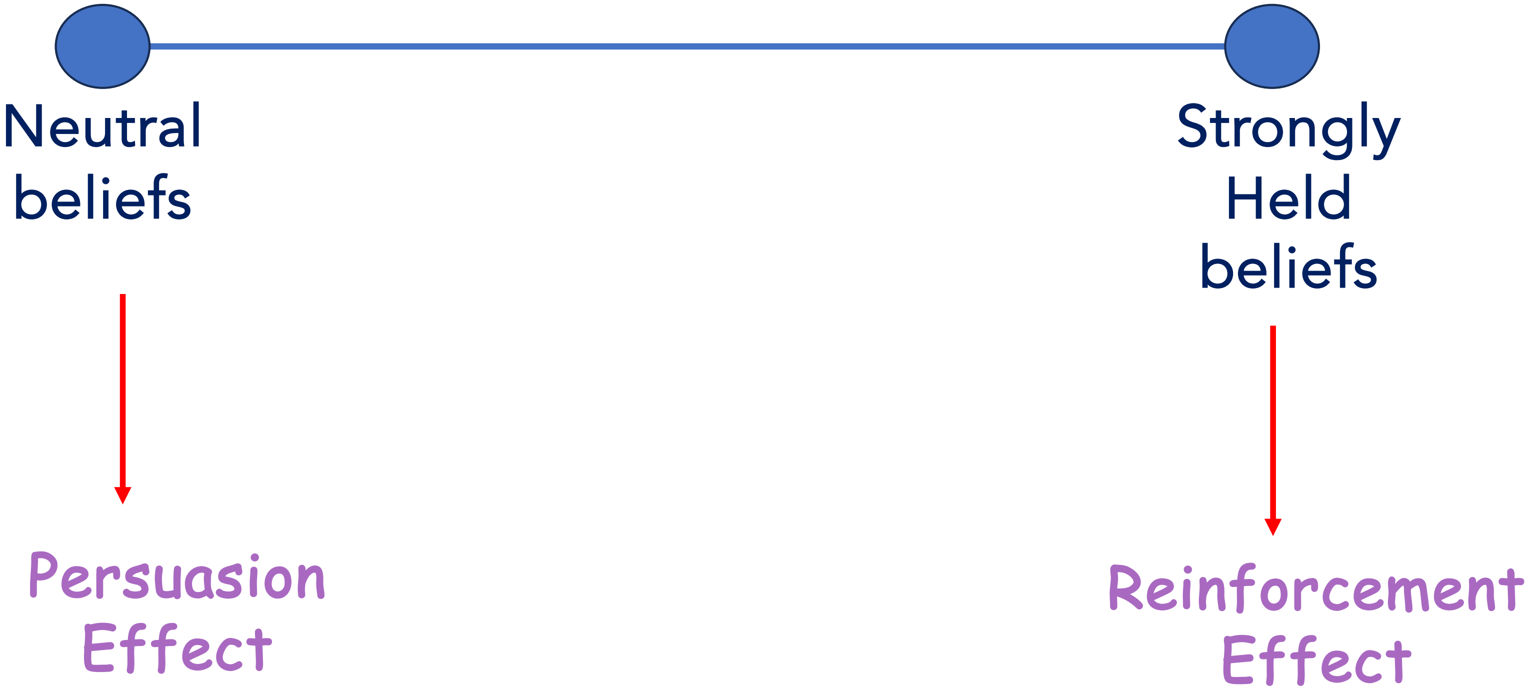
Kanizsa Triangle
- Optical illusion-

Sander van der Linden (2023).

Hasher, L. D. Goldstein, and T. Toppino (1977).

“HUMAN PERCEPTIONS OF TRUTH ”

Population - level



Neutral beliefs

Persuasion Effect

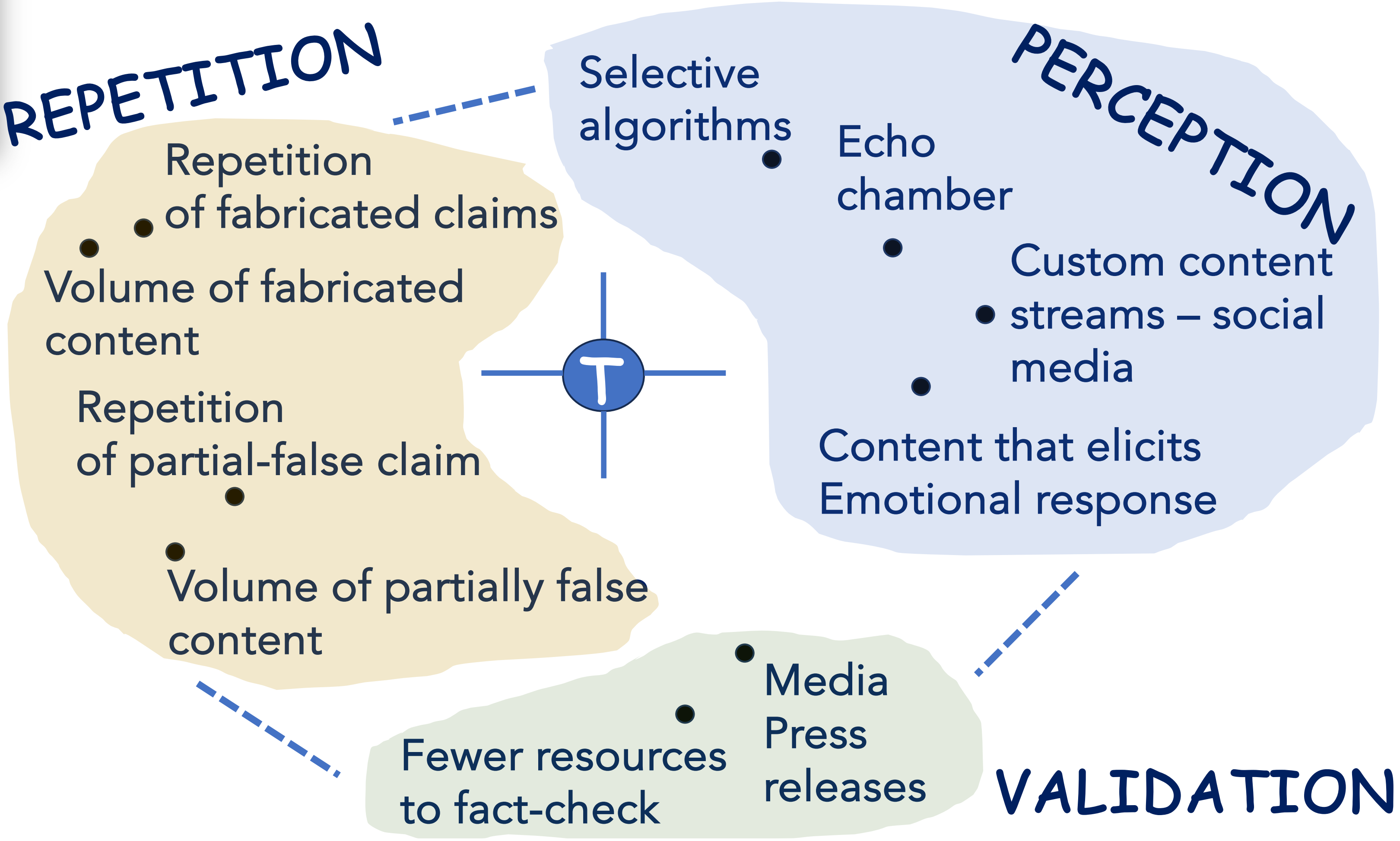
Strongly Held beliefs

Reinforcement Effect

Arkes, H. R., Boehm, L. E., & Xu, G. (1991).

THE SCIENCE OF MANIPULATION

"MANIPULATION OF THE TRUTH "



REPETITION

- Repetition of fabricated claims
- Volume of fabricated content
- Repetition of partial-false claim
- Volume of partially false content

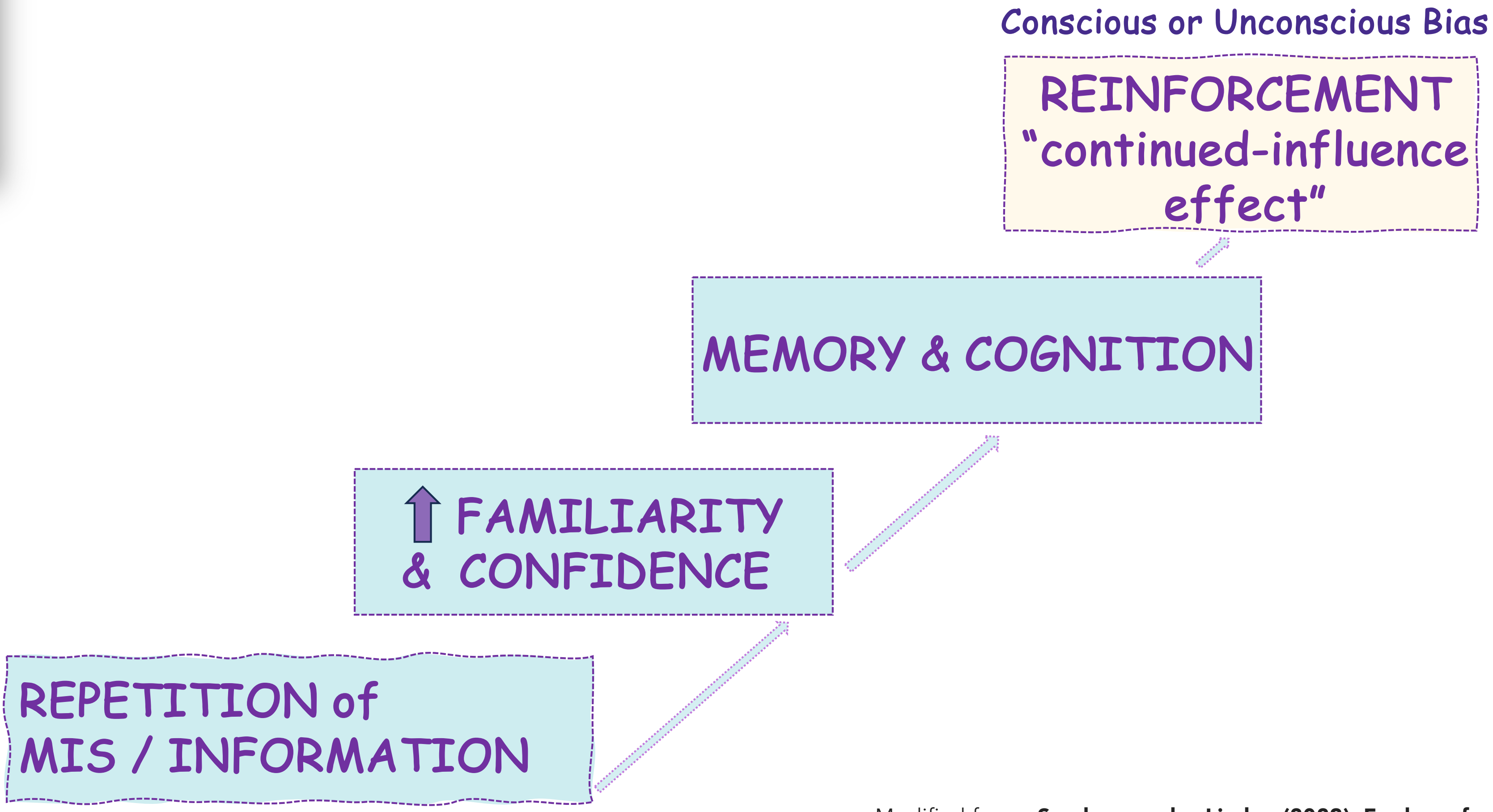
PERCEPTION

- Selective algorithms
- Echo chamber
- Custom content streams – social media
- Content that elicits Emotional response

VALIDATION

- Media Press releases
- Fewer resources to fact-check

“MANIPULATION OF THE TRUTH ”



REPETITION of MIS / INFORMATION

⬆ FAMILIARITY & CONFIDENCE

MEMORY & COGNITION

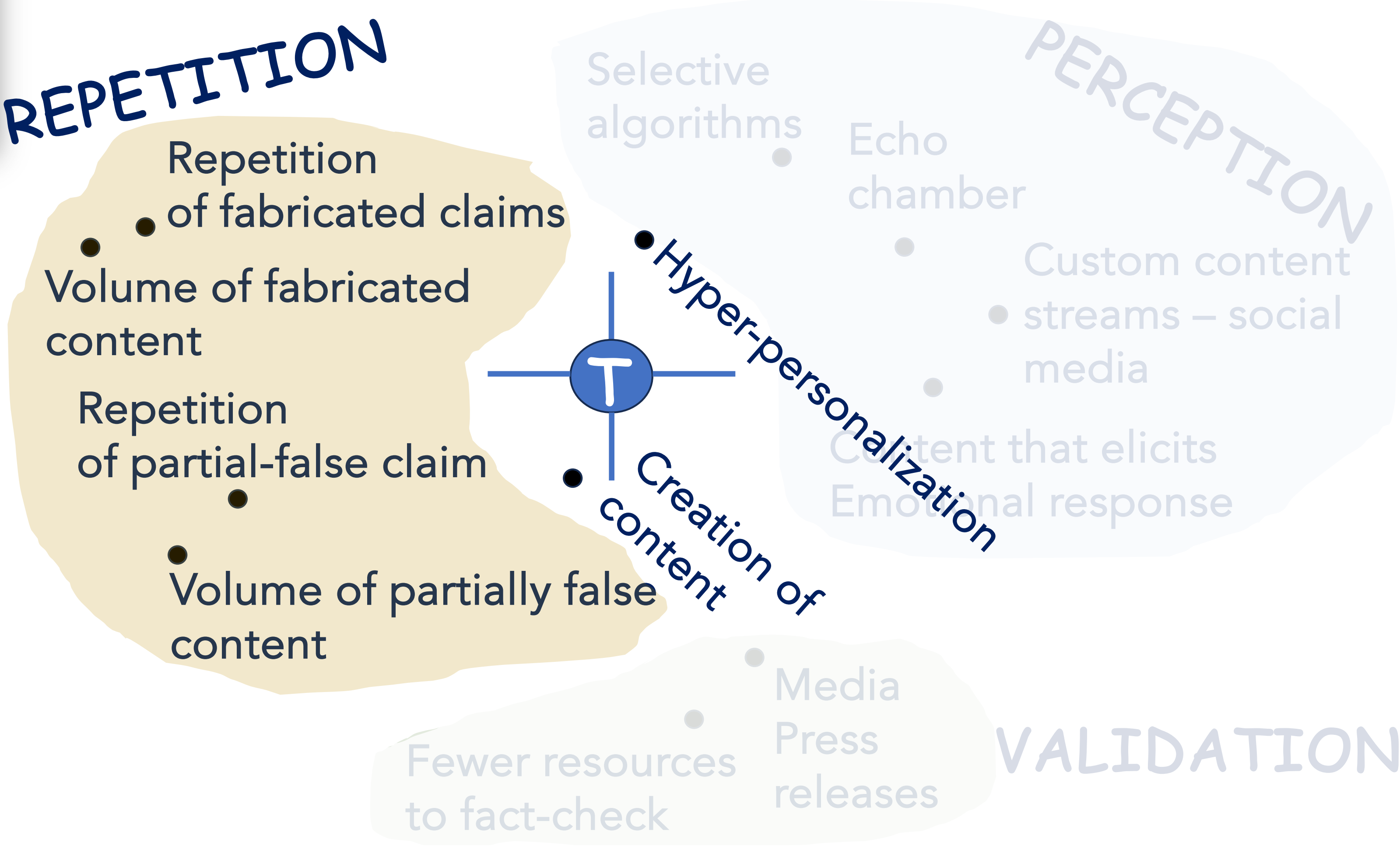
REINFORCEMENT “continued-influence effect”

Conscious or Unconscious Bias

Modified from: **Sander van der Linden (2023). Foolproof**: Why misinformation infects our minds and how to build immunity. WW Norton & Company.

WHAT IS THE ROLE OF AI?

“MANIPULATION OF THE TRUTH ”



REPETITION

- Repetition of fabricated claims
- Volume of fabricated content
- Repetition of partial-false claim
- Volume of partially false content

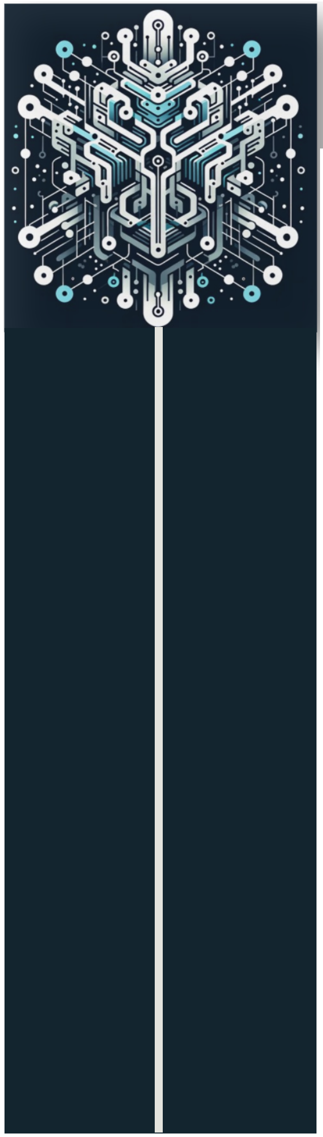
PERCEPTION

- Selective algorithms
- Echo chamber
- Custom content streams – social media
- Content that elicits Emotional response

VALIDATION

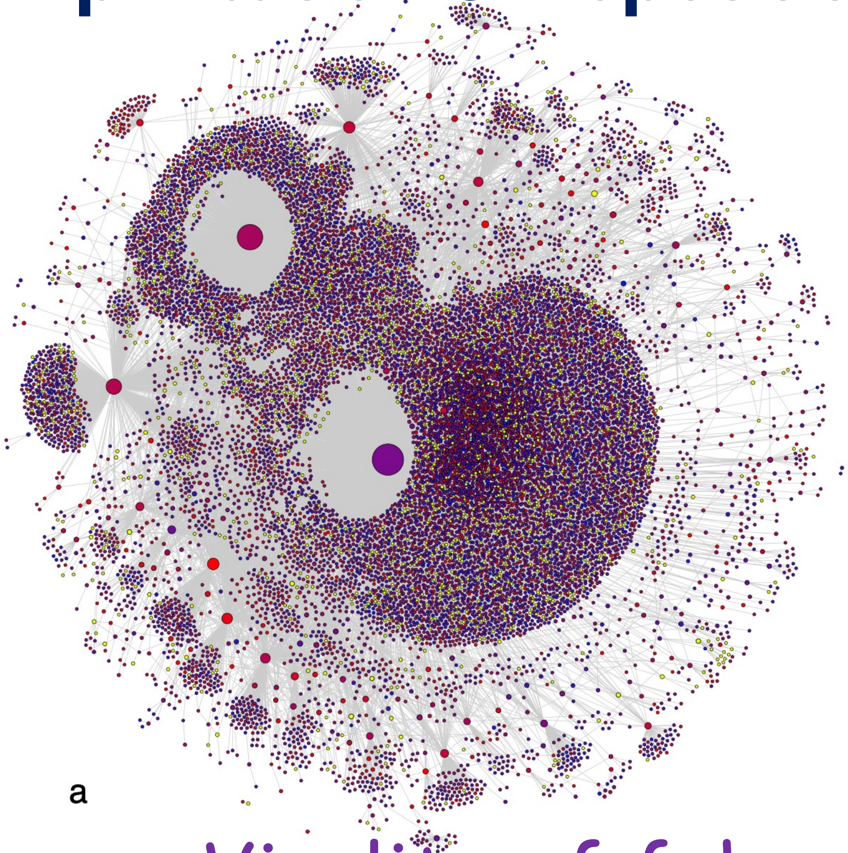
- Media Press releases
- Fewer resources to fact-check

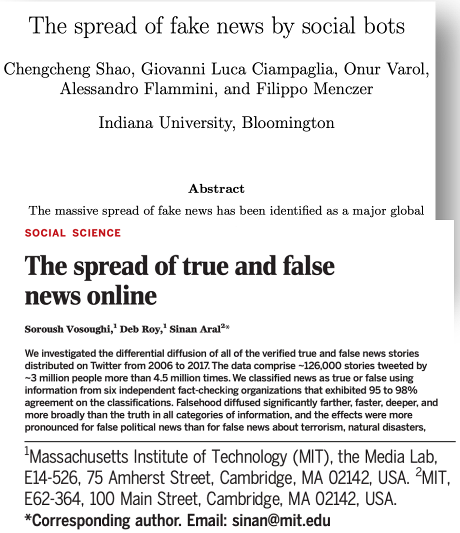
- Hyper-personalization
- Creation of content



"ROLE OF AI"

Amplification & Repetition of Falsehoods:





The spread of fake news by social bots

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol,
Alessandro Flammini, and Filippo Menczer

Indiana University, Bloomington

Abstract

The massive spread of fake news has been identified as a major global

SOCIAL SCIENCE

The spread of true and false news online

Soroush Vosoughi,[1] Deb Roy,[1] Sinan Aral[2]*

We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters,

[1]Massachusetts Institute of Technology (MIT), the Media Lab, E14-526, 75 Amherst Street, Cambridge, MA 02142, USA. [2]MIT, E62-364, 100 Main Street, Cambridge, MA 02142, USA.
*Corresponding author. Email: sinan@mit.edu

Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information,

Virality of fake news (INFODEMIC)

SPEED OF LIGHT

“ROLE OF AI“

Bias and the Training of Models:

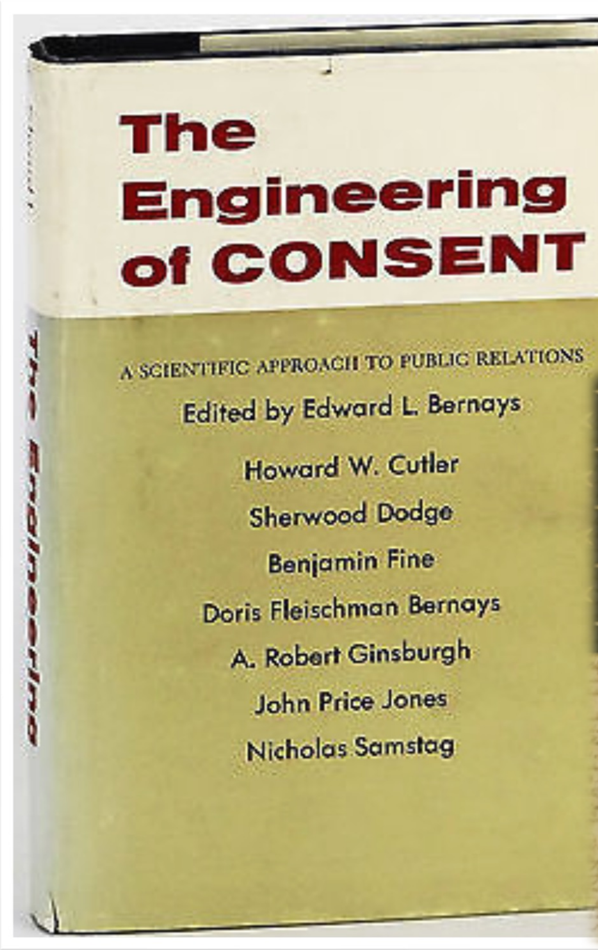
INFORMATION AGE

- **146 Zettabytes (or 146 trillion GB)** of information (ALL Data) – 2024
  - 23% growth rate (181 Zettabytes 2025)

- Equivalent to **9.4 million volumes** of the Encyclopedia Britannica per person (2,200 new volumes per year per person) !!

- Generative AI models are **trained on corpus of internet data**
- **AGNOSTIC** TECH - **CAN NOT distinguish “Fact” from “Fiction**”

ENGINEERING CONSENT - THE EDWARD BERNAYS EFFECT



The Engineering of CONSENT

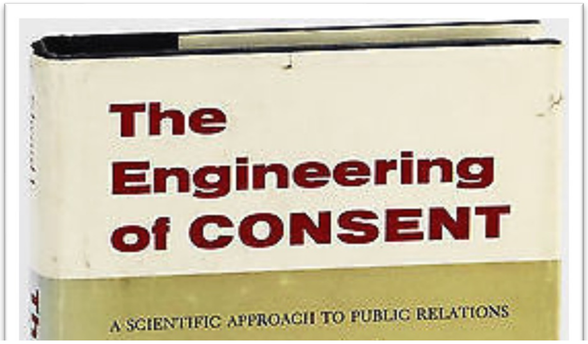
A SCIENTIFIC APPROACH TO PUBLIC RELATIONS

Edited by Edward L. Bernays

Howard W. Cutler
Sherwood Dodge
Benjamin Fine
Doris Fleischman Bernays
A. Robert Ginsburgh
John Price Jones
Nicholas Samstag

The Engineering

LUCKY STRIKE
40 MENTAL_FLOSS



The Engineering of CONSENT

A SCIENTIFIC APPROACH TO PUBLIC RELATIONS

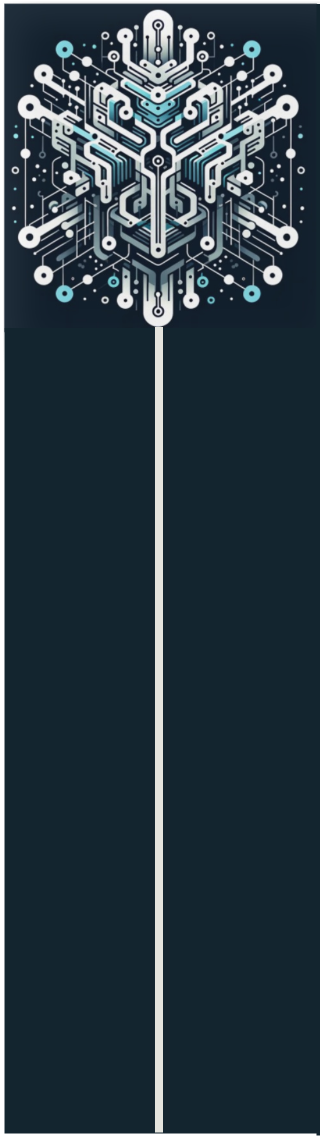
“The conscious and intelligent manipulation of the organized habits and opinions of the masses is an important element in a democratic society”

-Edward Bernays 1947

..Those that manufacture **consent** influence **power** – helping to support systems of **oppression**…

Mis(dis)-information

- Not new (centuries old)
- Difficult to debunk
- Informs perception and biases
- Persist



THE INTERSECTIONAL
INTERNET
RACE,SEX,CLASS, an
EDITED BY
Safiya Umoja Noble and
why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE
MANIPULATING the Message
How Powerful Forces Shape the News
CECIL ROSNER
Copyrighted Material
BEHIND THE HEADLINES
A History of Investigative Journalism in Canada
er
of a generation of rebels
work on the subject but
Starowicz, CBC Television