UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Moral Development in AI

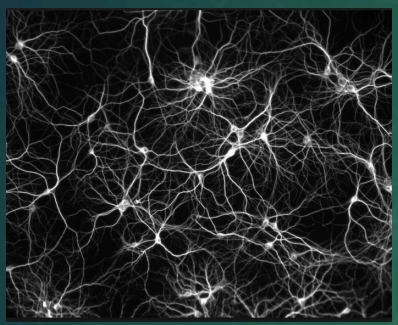
From psychometrics to psychographics

John Rust

http://www.psychometrics.cam.ac.uk
The Psychometrics Centre
Judge Business School
Trumpington Street

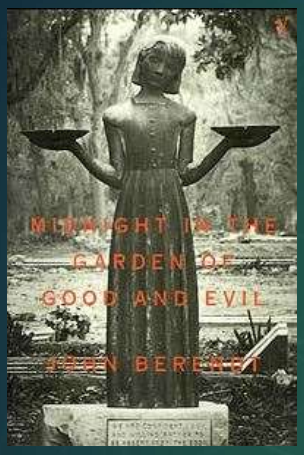
The disintegration of psychology

Neuropsychology



Psychometrics

20th century psychometrics

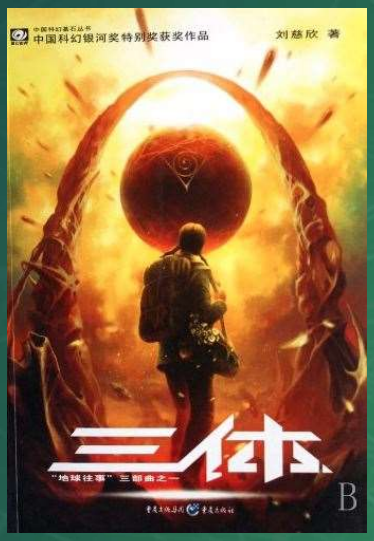


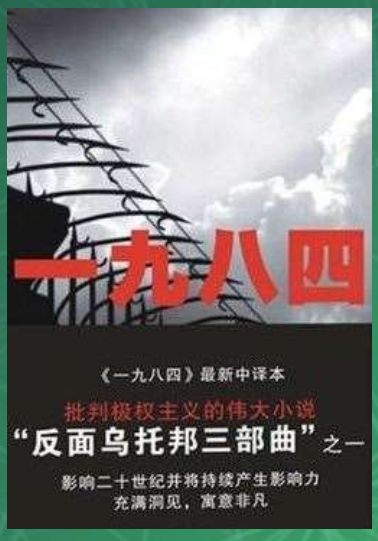
Samuel Taylor Coleridge (Table Talk, 1831)

If men could learn from history, what lessons it might teach us!

But passion and party blind our eyes, and the light which experience gives us … … is a lantern on the stern which shines only on the waves behind.

The 21st century. A clash of dystopias





The Three Body Problem: Chaotic Disruptors

- Innovations
  - 1: Psychographics
  - 2: Behavioural Science
  - 3: Cyber communication
  - 4: Machine Learning
  - 5: Artificial Intelligence
- Dystopic Visions
  - 1: 1984 (Orwell, 1948)
  - 2: Walden Two (Skinner, 1948)
  - 3: The Three Body Problem (Liu Cixin, 2007)
  - 4: I, Robot (Asimov, 1950)
  - 5: The Second Coming (Yeats, 1919)

Digital footprints are psychometric item pools

• **Private traits and attributes are predictable from digital records of human behavior** by M. Kosinski, D. Stillwell, T. Graepel, PNAS, 2013. (IF: 9.1) 2015.

–Online, the 21st most discussed paper of 2013 (Altmetric)

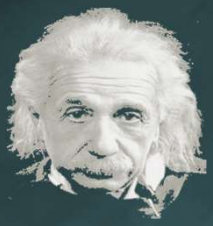
• **Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach** by A. Schwartz, J. …. M. Kosinski, D. Stillwell, M.E.P. Seligman, L.H. Ungar, PLoS ONE, 2013.

• **Computer-based personality judgments are more accurate than those made by humans** by W. Youyou, M. Kosinski, D. Stillwell, Proceedings of the National Academy of Sciences (PNAS), 2015,1

–Online, the 22nd most discussed paper of 2015 (Altmetric).

What we can predict from your digital footprint

BIG5 Personality

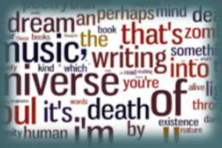


Intelligence

Life Satisfaction

Political Views

Religious Views



Use of Language

Use of addictive substances

Parents' relationship status

Interest/Profession

Relationship status

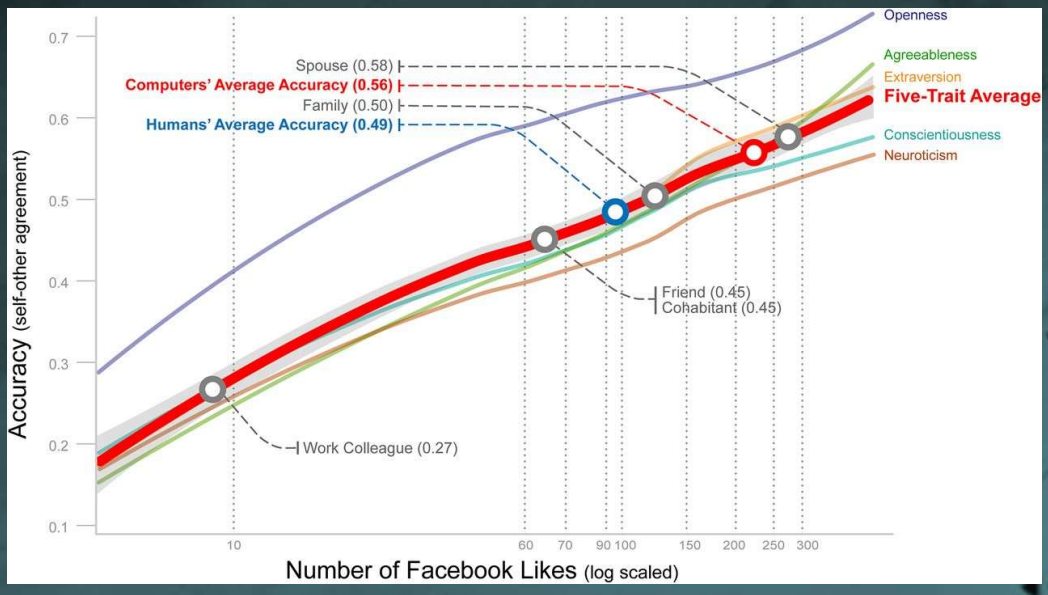
Sexual orientation

Ethnic Origin

Gender

Age

Wu, Y., Kosinski, M. & Stillwell, D. (2015) Computer-based personality judgements are more accurate than those made by humans. Proceedings of the National Academy of Sciences (PNAS)



IQ

High	Low
The Godfather	Jason Aldean
Mozart	Tyler Perry
Thunderstorms	Sephora
The Daily Show	Chiq
To Kill a Mockingbird	Bret Michaels
Lord of the Rings	Harley-Davidson
Science	Bebe

Extraversion

Introversion

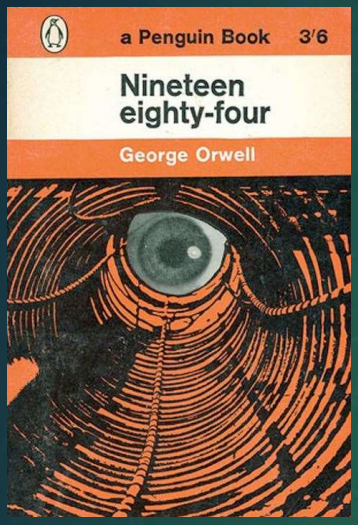
Emotionally stability

Neuroticism

Men

Women

The End of Privacy: 1984

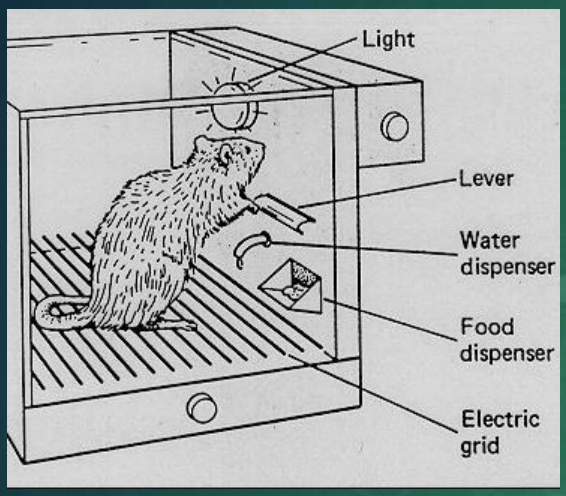


George Orwell, "1984" (1948)

"Always eyes watching you and the voice enveloping you. Asleep or awake, indoors or out of doors, in the bath or bed- no escape. Nothing was your own except the few cubic centimetres in your skull."

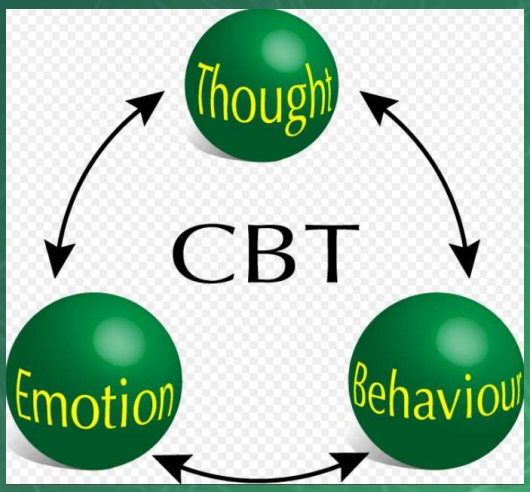
It was terribly dangerous to let your thoughts wander when you were in any public place or within range of a telescreen. The smallest thing could give you away. A nervous tic, an unconscious look of anxiety, a habit of muttering to yourself – anything that carried with it the suggestion of abnormality, of having something to hide. In any case, to wear an improper expression on your face (to look incredulous when a victory was announced, for example) was itself a punishable offense. There was even a word for it in Newspeak: facecrime, it was called. (1.5.65)

From Behaviourism to Brexit



Cognitive Behavioural Therapy (CBT)

- "CBT is based on the belief that thought disorders and maladaptive behaviours play a role in the development and maintenance of psychological disorders."
- Who decides what counts as a 'though disorder' or 'maladaptive behaviour'?
- E.g. Was 'Not being able to see the obvious advantages of Soviet Communism' a thought disorder?



The nudge initiative in UK politics

- "Improving decisions about Health, Wealth and Happiness" (Thaler and Sunstein, 2008)
- Soft paternalism: to encourage people to remedy 'incorrect behaviour' by covert means: E.g.
  - Not saving enough for retirement
  - Eating too much junk food
  - Smoking
  - Not getting vaccinated
  - Excessive use of energy
  - Not recycling your rubbish

The UK Behavioural Insights Team

- House of Lords Report (2011) "(Nudges) prompt choices without getting people to consider their options consciously, and therefore do not include openly persuasive interventions such as media campaigns and the straightforward provision of information"
- Cabinet Office (2014) "The Behavioural Insights Team, often called the 'Nudge Unit', applies insights from academic research in behavioural economics and psychology to public policy and services."

Those at risk

- Health
  - Eating disorders
  - Lifestyle
  - Depression
- Behaviour
  - Dangerous driving
  - Substance abuse
  - Bullying

Those who place others at risk

- Criminals
- Terrorists
- Cyber-criminals
- Sex offenders
- Fraudsters
- 'Insider Threat'
- Political extremists

Getting it all joined up

Records of health, education, crime, tax, credit, spending, travel, work, genealogy, internet and mobile use, location, webcams

All along the digital superhighway, you will be nudged

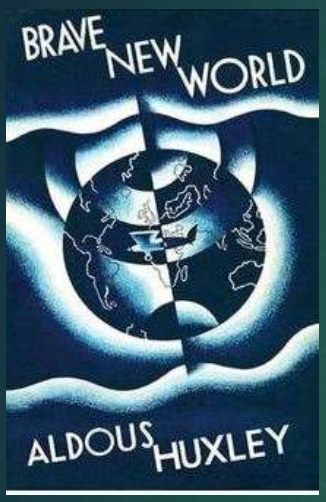
- In the interests of your health, welfare and the environment.
- And to protect society, should you in any way present a threat
- "If you have nothing to hide you have nothing to fear"
  (William Hague, UK Foreign Minister, 2013)
- *如果你没有什么可隐瞒的，那么你将无所畏惧*

Behavioural Science in the online environment

- **Online prediction**
- Your digital footprint consists of items
- that can be analysed psychometrically
- We know what you are likely to do
- We will send you special messages
- ***MICRO-TARGETED*** just for you

- **Online control**
- We decide what you see, and what you don't see. This determines what you and people like you are likely to say and do.
- You and your friends live in a
- ***SILO / ECHO-CHAMBER***

Behavioural Science: The Brave New World



Aldous Huxley, "Brave New World" (1931)

B. F. Skinner, "Walden Two" (1948)

"Utopian society is possible, but only through behavioural engineering. Human beings are sometimes selfish, greedy and mean. Human nature must be changed, engineered so that people are non-competitive, happy and harmonious. Positive rewards can change both outward behaviour and inner motives. People can be 'conditioned' to live together in peace and harmony."

Classical communication

"There is no need for cyberspace. All internet related things can be handled with existing conceptual categories. The internet has nothing to do with technology. It is just a person talking to another person, using a different technology". *(Souls writing on the net, Giancarlo Livraghi, 2001)*

Modern communication "The Medium is the Message"

- Cyberspace is an interactive domain made up of digital networks that is used to store, modify and communicate information. It includes the internet, but also the other information systems that support our businesses, infrastructure and services". *(UK Cyber Security Strategy, 2011)*
- "For now we see through a glass darkly; but then face to face. Now I know in part, but the I shall know even as also I am known." (*St Paul, Letter to the Corinthians 13: 12-13*)

Cyberspace as The Dark Forest

Cyberspace as The Dark Forest

Cyberspace as The Dark Forest

Cyberspace as The Dark Forest

Cyberspace as The Dark Forest

Living with Chaos: The Three Body Problem

- **The Second Coming"**
  William Butler Yeats

"Turning and turning in the widening gyre
The falcon cannot hear the falconer;
Things fall apart; the centre cannot hold;
Mere anarchy is loosed upon the world,
The blood-dimmed tide is loosed, and everywhere
The ceremony of innocence is drowned;
The best lack all conviction, while the worst
Are full of passionate intensity.
Surely some revelation is at hand;
Surely the Second Coming is at hand.

From Neural Networks to Robots

- Our digital footprints (Avatars)
- Bots (Cyber Robots),
- Machine Learning,
- Genetic algorithms,
- Artificial Intelligence (AI)

Explaining Artificial Intelligence (XAI)

"As a society, we must be able to look into the 'black box' of big data analytics in order to ensure that any particular analytics application can be safely installed and will benefit us all".

European Data Protection Supervisor Opinion, July 2015
"Meeting the Challenges of Big Data"

What's inside the Black Box?

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

Insurance

Millions of Britons paying 'ethnic minority penalty' for car insurance

Drivers in areas with high density of minority ethnic households are each charged up to £450 a year more, says study

ALGORITHMIC INEQUITY 7/8/15 1:33 PM

Google showed women ads for lower-paying jobs

Supreme Court vs. Neighborhood Segregation

In a surprising move on Thursday, the United States' highest court ruled that policies even inadvertently relegating minorities to poor areas violate the Fair Housing Act.

Machine Bias

The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review

A blot on the profession

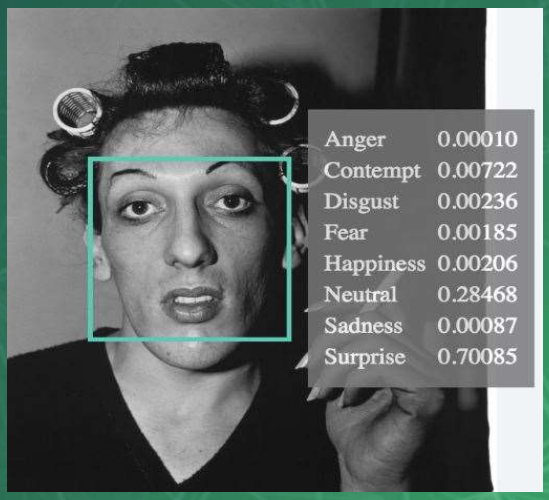
Discrimination in medicine against women and members of ethnic minorities has long been suspected,[1] but it has now been proved. St George's Hospital Medical School has been found guilty by the Commission for Racial Equality of practising racial and sexual discrimination in its admissions policy.[2] The commission decided not to serve a non-

reassuring as it
other schools.
The commis
about this par
schools can avo
where a compu

How AI understands human psychology

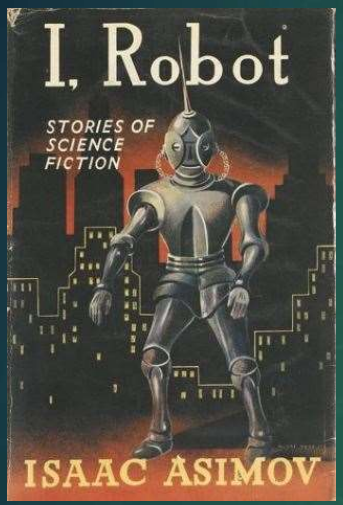
Google Deepmind (Alpha Go)

Microsoft (Project Oxford)



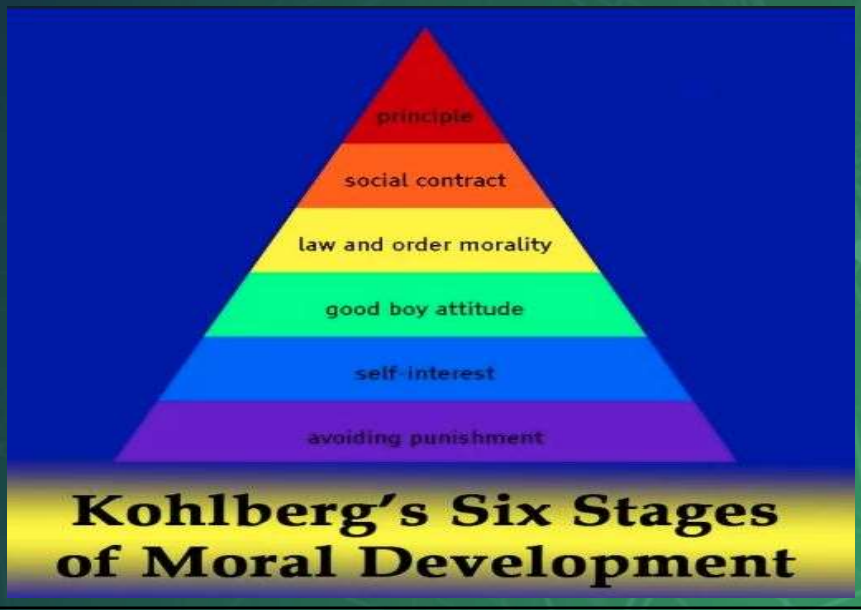
Moral development

Asimov's vision: the well-behaved robot



- 1: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- 2: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- 3: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
- 0: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

The Psychology of Moral Development



Ethics vs. Law

Law	Ethics
Formal, written document	Unwritten principles
Interpreted by courts	Interpreted by each individual
Established by legislatures	Presented by philosophers, religious, professional groups
Applicable to everyone	Personal choice
Priority decided by court	Priority determined by individual
Court makes final decision	No external decision maker
Enforceable by police and courts	Limited enforcement

Can machines behave ethically?

Humans

- Thinking
- Experiencing Emotions
- Theory of Mind
- Moral Development (Kohlberg)
  - 1. Self-interest (avoid punishment)
  - 2. 'Good and Bad' morality
  - 3. Law and Order
- 4. Principled Ethics

Machines

- Calculating
- Identifying individual differences
- Recognising emotions
- Specifying terms of interaction
  - Maximize Click-through and ROI
  - Be nice to children
  - Laws of Robotics
- Unknown

Moral Development in intelligent machines

Reward and Punishment

Imitation

Ethics?

What if principled conscience fails to develop?

- Psychopathic personality disorder
- Moral development 'frozen' at early stage stage
- Without conscience (Hare)
  - No feeling of remorse or shame
- The Mask of Sanity (Checkley)
  - Mimic normal person
  - Superficially charming
  - Do not experience emotions
- No understanding of 'right' and 'wrong'

Our robots are our future (too)

Urgent need for AI regulation