AI: Law and Policy Intersections

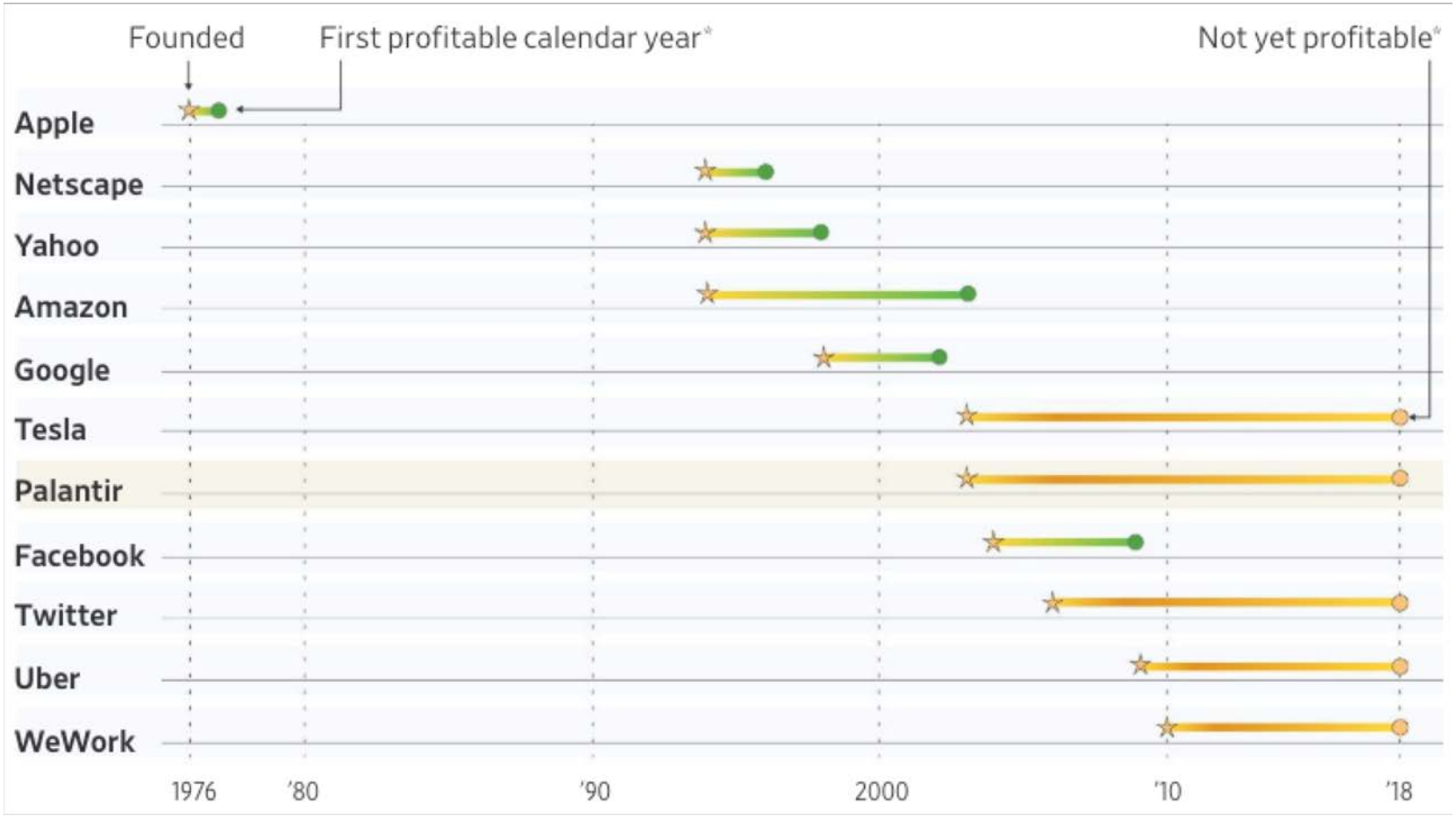
Chris Jay Hoofnagle
UC Berkeley School of Law, School of Information
Nov. 21, 2018

Roadmap

- **Level setting**
- AI and GDPR
- AI and military applications
- AI and quantum computing

Tech/political/economic landscape

- Affordances
- Winner-take-all effect because of ”data advantage”
  - Halevy et al (2009)
- EU protectionism versus US/China industrial policy?
  - Relatively loose data protection in China/US may advantage ML development
  - Presumably systems can be trained outside EU, but then deployed wrt Europeans
    - Many tools are open source; democratization of tools as a shaping force of AI
- Stakes high—driving hype
  - Uber will fail without autonomous driving
  - Google desperate for post-PageRank innovation



Wall Street Journal, Nov. 12, 2018

THE NEW YORKER

A REPORTER AT LARGE OCTOBER 22, 2018 ISSUE

DID UBER STEAL GOOGLE'S INTELLECTUAL PROPERTY?

Silicon Valley was built on job-hopping. But when a leader of Google's self-driving-car unit joined Uber, Google filed suit. Now the Feds are on the case.

By Charles Duhigg

Page was adamant. According to internal Google e-mails, he ordered executives to "make Anthony rich if Chauffeur succeeds." Two months later, Google bought 510 Systems for twenty-two million dollars. It also purchased Anthony's Robots; in return, Levandowski was guaranteed a future payment tied to the total value of Project Chauffeur. Google agreed to give him a claim on ten per cent of the division's eventual worth—a kind of shadow equity that would vest in four years. The stake eventually paid him more than a hundred and twenty million dollars, one of the largest such payouts in Google's history.

One day in 2011, a Google executive named Isaac Taylor learned that, while he was on paternity leave, Levandowski had modified the cars' software so that he could take them on otherwise forbidden routes. A Google executive recalls witnessing Taylor and Levandowski shouting at each other. Levandowski told Taylor that the only way to show him why his approach was necessary was to take a ride together. The men, both still furious, jumped into a self-driving Prius and headed off.

The car went onto a freeway, where it travelled past an on-ramp. According to people with knowledge of events that day, the Prius accidentally boxed in another vehicle, a Camry. A human driver could easily have handled the situation by slowing down and letting the Camry merge into traffic, but Google's software wasn't prepared for this scenario. The cars continued speeding down the freeway side by side. The Camry's driver jerked his car onto the right shoulder. Then, apparently trying to avoid a guardrail, he veered to the left; the Camry pinwheeled across the freeway and into the median. Levandowski, who was acting as the safety driver, swerved hard to avoid colliding with the Camry, causing Taylor to injure his spine so severely that he eventually required multiple surgeries.

Since 2014, California regulations have required companies to report any instance in which a self-driving vehicle is "in any manner involved in a collision originating from the operation of the autonomous vehicle on a public road that resulted in the damage of property or in bodily injury or death." The Camry accident occurred three years before this regulation was passed; since the rule went into effect, Google has reported thirty-six additional accidents. If Google is still failing to report accidents in which its cars did not hit other vehicles, then there may have been more undocumented incidents. "There's lots of times something happened because one of our cars drove erratically but never hit anyone," a former senior Google executive told me. Google cars sometimes stopped suddenly, including at intersections, causing other cars to swerve. (A spokesperson for Google declined to discuss the company's reporting policies.)

Positive vision

- Augmented intelligence/reality
- Feedback on behavior, insight into inner state
- Sensemaking (eyeglasses > affective states)
- Provision of context, even hedge against “fake news”
- Scut work performed by machine
- Even creative work enhanced by machines
  - Computers generate musical options
- Advances in optimization in many industries

Roadmap

- Level setting
- **AI and GDPR**
- AI and military applications
- AI and quantum computing

ML as a new kind of discipline

- What are the institutional, substantive, and procedural protections necessary for this new form of social ordering?
- How could the GDPR be made ML-friendly?
  - Must incorporate the global strategic posture of Asia and the US
  - Different domains of ML will need different approaches
  - Affordances matter
    - Tools are open source and freely available
    - Mega data firms
    - ML likely to exist in personal devices
- What self-regulation is realistic, effective?

Privacy's goals

- Privacy is an instrumental value; often articulated as a terminal one
- Minimization
- Procedural protections
- At the highest level, minimization & procedures are intended to blunt the power of decision makers (give users rights & impose responsibilities on data holders)

GDPR & ML

- Ex ante dominant
- Minimization & use limitations
- Consent can't be the catch-all exception
- Assigns ad optimization to "high risk" data processing
- Article 15: "meaningful information about the logic involved…"
- Article 22: right against being "subject to a decision based solely on automated processing, including profiling…"

THE EUROPEAN UNION GENERAL DATA PROTECTION REGULATION: WHAT IT IS AND WHAT IT MEANS

Chris Jay Hoofnagle, Bart van der Sloot, Frederik Zuiderveen Borgesius[1]

choofnagle [at] berkeley.edu, B.vdrsloot [at] uvt.nl,
Frederik.Zuiderveen.Borgesius [at] vub.ac.be

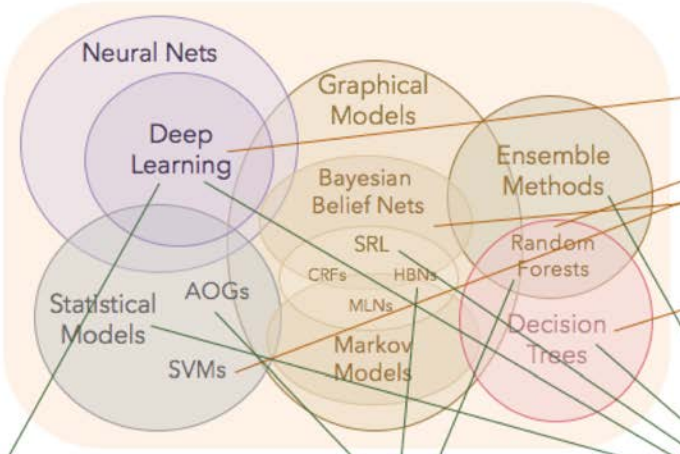
Abstract – This article introduces U.S. lawyers and academics to the normative foundations, attributes, and strategic approach to regulating personal data advanced by the European Union's General Data Protection Regulation ("GDPR"). We explain the genesis of the GDPR, which is best understood as an extension and refinement of existing requirements imposed by the 1995 Data Protection Directive; describe the GDPR's approach and provisions; and make predictions about the GDPR's short and medium-term implications. The GDPR is the most consequential regulatory development in information policy in a generation. The GDPR brings personal data into a detailed and protective regulatory regime, which will influence personal data usage worldwide. Understood properly, the GDPR encourages firms to develop information governance frameworks, to in-house data use, and to keep humans in the loop in decision making. Companies with direct relationships with consumers have strategic advantages under the GDPR, compared to third party advertising firms on the internet. To reach these objectives, the GDPR uses big sticks, structural elements that make proving violations easier, but only a few carrots. The GDPR will complicate and restrain some information-intensive business models. But the GDPR will also enable approaches previously impossible under less-protective approaches.

Performance vs. Explainability

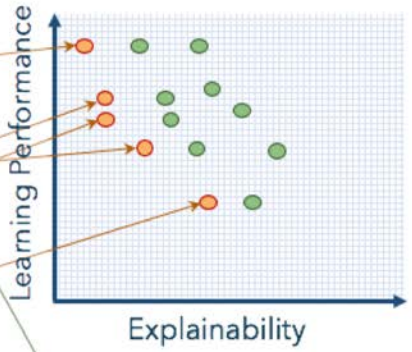
New Approach

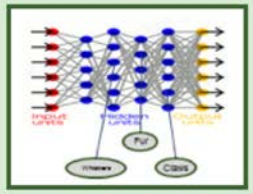
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



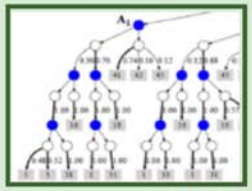
Explainability (notional)





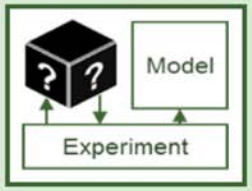
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

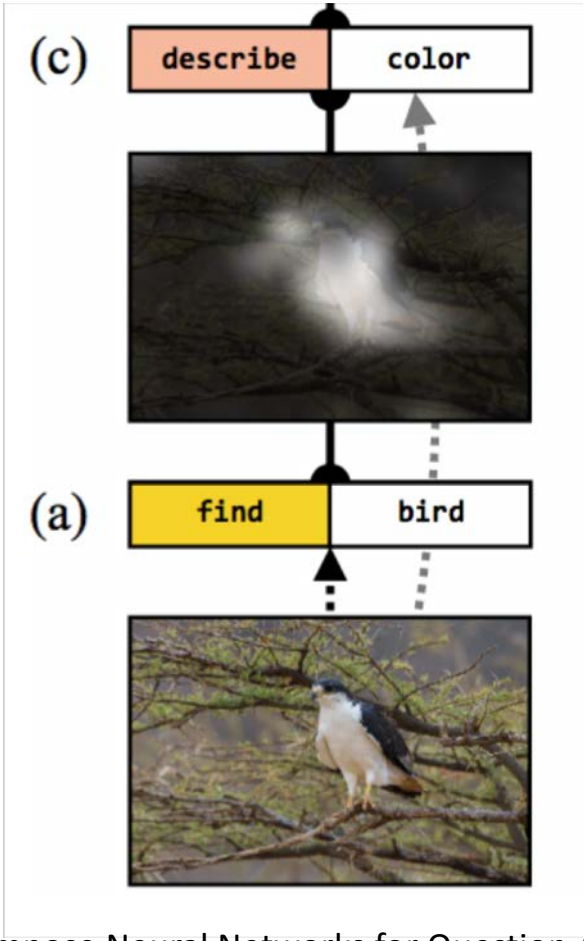


Model Induction

Techniques to infer an explainable model from any model as a black box

Explainability challenges

- The “Alien Intelligence” metaphor
- Complexity
- Deliberate non-transparency
- Emergent outcomes
- Lack of justification
- Counterfactuality



Andreas et al, Learning to Compose Neural Networks for Question Answering

Problems in algorithmic “fairness”

- Confusion over instrumental and terminal goals
  - An autonomous weapons system could kill child combatants lawfully; is this “fair”
  - Is the point to be fair or to avoid unfairness?
  - Emphasis on procedural protections (much like privacy) rather than substance
- Semantic sleight of hand
  - Algorithmic fairness could be “accountable” in two senses:
    - Making an accounting of an action
    - Making people/institutions responsible for actions
- Still need to decide what “fairness” is being promoted

Mechanics of fairness

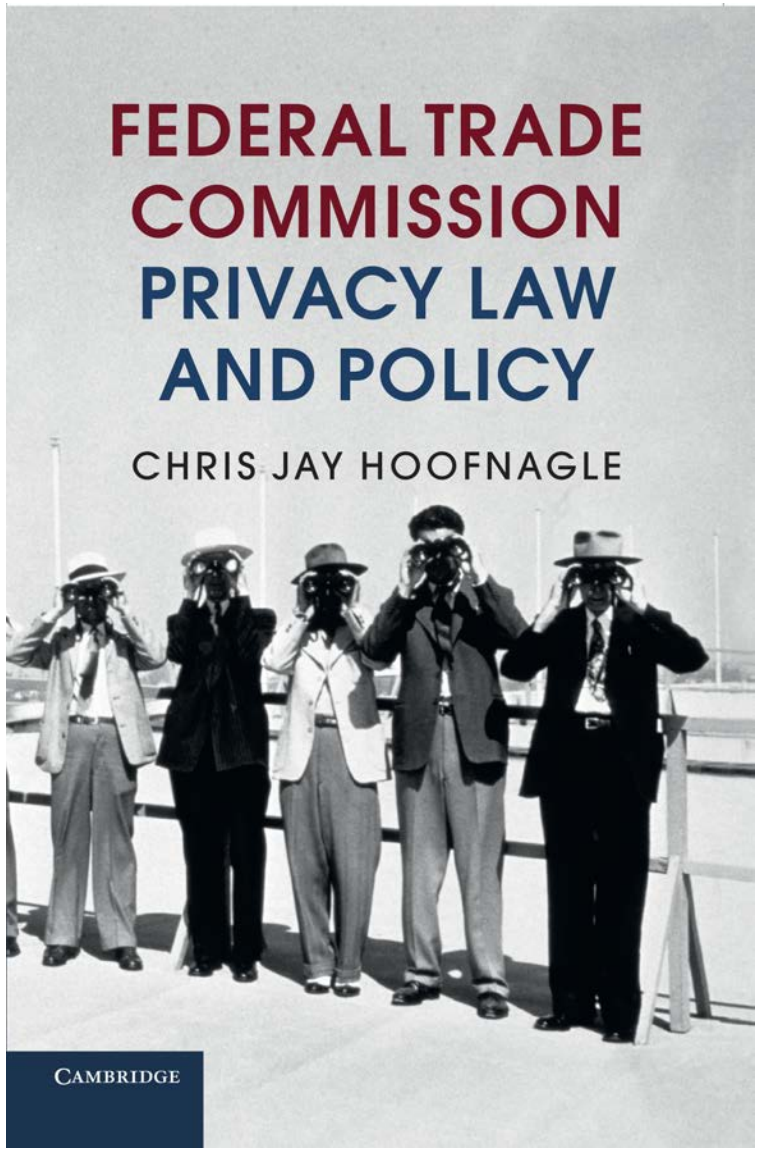
- Insight is needed into
  - Data inputs
  - Algorithms
  - Decisions made
- Each of these areas can have nuanced biases
  - In data selection, cleaning, etc
  - In institutional practices & assumptions
  - In technical limitations that shape analysis
  - Could be emergent forms of bias

Subtle design choices have real consequences

- Let's say you wanted to use NLP to detect ISIS "travelers" on Twitter. Methods matter–
- Techniques to reduce the problem space damage context
  - Stopwords often eliminate negatives, such as "wouldn't," as in "I wouldn't travel to Istanbul to join ISIS"
  - Vectorizing can reduce context
  - Size of N-grams matters
  - Tradeoffs between stemming and lemmas may result in choosing the more inaccurate stemming instead of resource-intensive, contextual lemma.
    - E.g. meanness and meaning become mean under stemming.
- Techniques to optimize can focus on false positives or false negatives

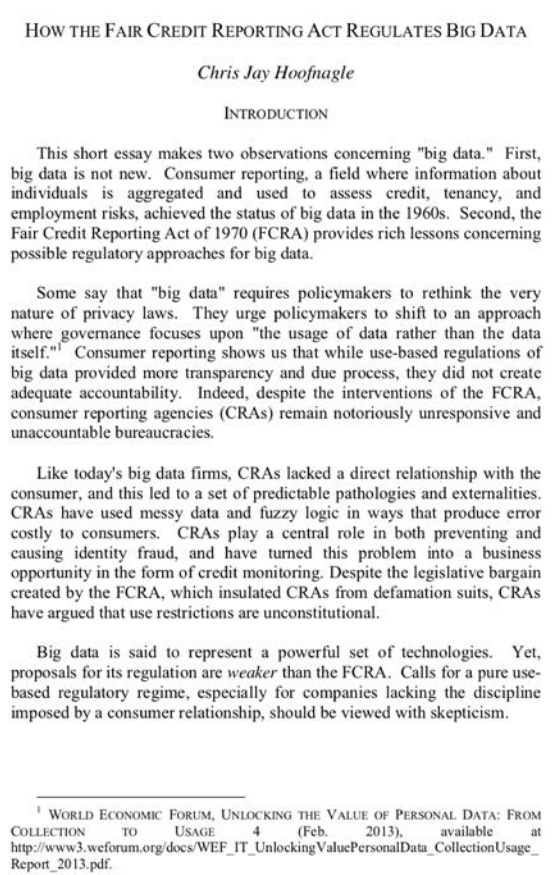
Expert regulatory institutions

- Vest regulators with authority, expertise to address changing situations
- Many downsides…



Fair credit reporting act

- Credit reporting (CRAs) had ”big data” status in 1960s
- Expert systems ”AI”
- 1970 Legislative solution was a performance-based standard:
  - CRAs shielded from defamation liability if they adopted techniques to ensure “maximum possible accuracy”
  - Terminal goal: accuracy
  - Instrument: performance standard + procedural rights
  - Ex post approach



HOW THE FAIR CREDIT REPORTING ACT REGULATES BIG DATA

Chris Jay Hoofnagle

INTRODUCTION

This short essay makes two observations concerning "big data." First, big data is not new. Consumer reporting, a field where information about individuals is aggregated and used to assess credit, tenancy, and employment risks, achieved the status of big data in the 1960s. Second, the Fair Credit Reporting Act of 1970 (FCRA) provides rich lessons concerning possible regulatory approaches for big data.

Some say that "big data" requires policymakers to rethink the very nature of privacy laws. They urge policymakers to shift to an approach where governance focuses upon "the usage of data rather than the data itself."[1] Consumer reporting shows us that while use-based regulations of big data provided more transparency and due process, they did not create adequate accountability. Indeed, despite the interventions of the FCRA, consumer reporting agencies (CRAs) remain notoriously unresponsive and unaccountable bureaucracies.

Like today's big data firms, CRAs lacked a direct relationship with the consumer, and this led to a set of predictable pathologies and externalities. CRAs have used messy data and fuzzy logic in ways that produce error costly to consumers. CRAs play a central role in both preventing and causing identity fraud, and have turned this problem into a business opportunity in the form of credit monitoring. Despite the legislative bargain created by the FCRA, which insulated CRAs from defamation suits, CRAs have argued that use restrictions are unconstitutional.

Big data is said to represent a powerful set of technologies. Yet, proposals for its regulation are *weaker* than the FCRA. Calls for a pure use-based regulatory regime, especially for companies lacking the discipline imposed by a consumer relationship, should be viewed with skepticism.

[1] WORLD ECONOMIC FORUM, UNLOCKING THE VALUE OF PERSONAL DATA: FROM COLLECTION TO USAGE 4 (Feb. 2013), available at http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf.

FCRA can deal with emergent problems

- Judy **Thomas** sued TransUnion for regularly mixing her report with a Jud**ith Upton**.
- Thomas' SSN was on digit different from Upton's + they shared "Jud" in the first name
- CRAs expert systems classified women by their first names!

Los Angeles Times

Jury Awards $5.3 Million for Credit Report Errors

July 31, 2002 | KATHY M. KRISTOF | TIMES STAFF WRITER

An Oregon woman who fought for six years to clear erroneous items from her credit report was awarded $5.3 million Monday by a federal jury in Oregon.

The verdict against Chicago-based credit reporting firm Trans Union was the largest amount ever awarded for violations of the Fair Credit Reporting Act, which requires credit reporting companies to keep accurate records and promptly correct mistakes, consumer attorneys said Tuesday.

From procedure to substance

- Best data & tools are now in the private sector
  - Academics need more industry partnerships…
- Shoshana Zuboff: advances in ML are shifting the "division of learning" in society
  - Who knows?
  - Who decides?
  - Who decides who decides?
    - THE AGE OF SURVEILLANCE CAPITALISM (2018)
- Cf. "freedom to observe," "freedom to learn"
  - Private decision space

Personality judgments

Our personality traits and political predispositions are predictable from the “likes” we give away free on Facebook

…people’s personalities can be predicted automatically and without involving human social-cognitive skills.

PNAS

Computer-based personality judgments are more accurate than those made by humans

Wu Youyou[a,1,2], Michal Kosinski[b,1], and David Stillwell[a]

[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and [b]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

Judging others’ personalities is an essential skill in successful social living, as personality is a key driver behind people’s interactions, behaviors, and emotions. Although accurate personality judgments stem from social-cognitive skills, developments in machine learning show that computer models can also make valid judgments. This study compares the accuracy of human and computer-based personality judgments, using a sample of 86,220 volunteers who completed a 100-item personality questionnaire. We show that (*i*) computer predictions based on a generic digital footprint (Facebook Likes) are more accurate ($r = 0.56$) than those made by the participants’ Facebook friends using a personality questionnaire ($r = 0.49$); (*ii*) computer models show higher interjudge agreement; and (*iii*) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores. Computers outpacing humans in personality judgment presents significant opportunities and challenges in the areas of psychological assessment, marketing, and privacy.

personality judgment | social media | computational social science | artificial intelligence | big data

Perceiving and judging other people’s personality traits is an essential component of social living (1, 2). People use personality judgments to make day-to-day decisions and long-term plans in their personal and professional lives, such as whom to befriend, marry, trust, hire, or elect as president (3). The more accurate the judgment, the better the decision (2, 4, 5). Previous research has shown that people are fairly good at judging each other’s personalities (6–8); for example, even complete strangers can make valid personality judgments after watching a short video presenting a sample of behavior (9, 10).

Although it is typically believed that accurate personality perceptions stem from social-cognitive skills of the human brain, recent developments in machine learning and statistics show that computer models are also capable of making valid personality judgments by using digital records of human behavior (11–13). However, the comparative accuracy of computer and human judgments remains unknown; this study addresses this gap.

Personality traits, like many other psychological dimensions, are latent and cannot be measured directly; various perspectives exist regarding the evaluation criteria of judgmental accuracy (3, 5). We adopted the realistic approach, which assumes that personality traits represent real individual characteristics, and the accuracy of personality judgments may be benchmarked using three key criteria: self-other agreement, interjudge agreement, and external validity (1, 5, 7). We apply those benchmarks to a sample of 86,220 volunteers,* who filled in the 100-item International Personality Item Pool (IPIP) Five-Factor Model of personality (14) questionnaire (15), measuring traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Computer-based personality judgments, based on Facebook Likes, were obtained for 70,520 participants. Likes were previously shown to successfully predict personality and other psychological traits (11). We used LASSO (Least Absolute Shrinkage and Selection Operator) linear regressions (16) with 10-fold cross-validations, so that judgments for each participant were made using models developed on a different subsample of participants and their Likes. Likes are used by Facebook users to express positive association with online and offline objects, such as products, activities, sports, musicians, books, restaurants, or websites. Given the variety of objects, subjects, brands, and people that can be liked and the number of Facebook users (>1.3 billion), Likes represent one of the most generic kinds of digital footprint. For instance, liking a brand or a product offers a proxy for consumer preferences and purchasing behavior; music-related Likes reveal music taste; and liked websites allow for approximating web browsing behavior. Consequently, Like-based models offer a good proxy of what could be achieved based on a wide range of other digital footprints such as web browsing logs, web search queries, or purchase records (11).

Human personality judgments were obtained from the participants’ Facebook friends, who were asked to describe a given participant using a 10-item version of the IPIP personality measure. To compute self-other agreement and external validity, we used a sample of 17,622 participants judged by one friend; to calculate interjudge agreement, we used a sample of 14,410 participants

Significance

This study compares the accuracy of personality judgment—a ubiquitous and important social-cognitive activity—between computer models and humans. Using several criteria, we show that computers’ judgments of people’s personalities based on their digital footprints are more accurate and valid than judgments made by their close others or acquaintances (friends, family, spouse, colleagues, etc.). Our findings highlight that people’s personalities can be predicted automatically and without involving human social-cognitive skills.

Author contributions: W.Y. and M.K. designed research; W.Y., M.K., and D.S. performed research; W.Y. and M.K. contributed new reagents/analytic tools; W.Y. and M.K. analyzed data; and W.Y., M.K., and D.S. wrote the paper.

Conflict of interest statement: D.S. received revenue as the owner of the myPersonality Facebook application.

This article is a PNAS Direct Submission. D.F. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

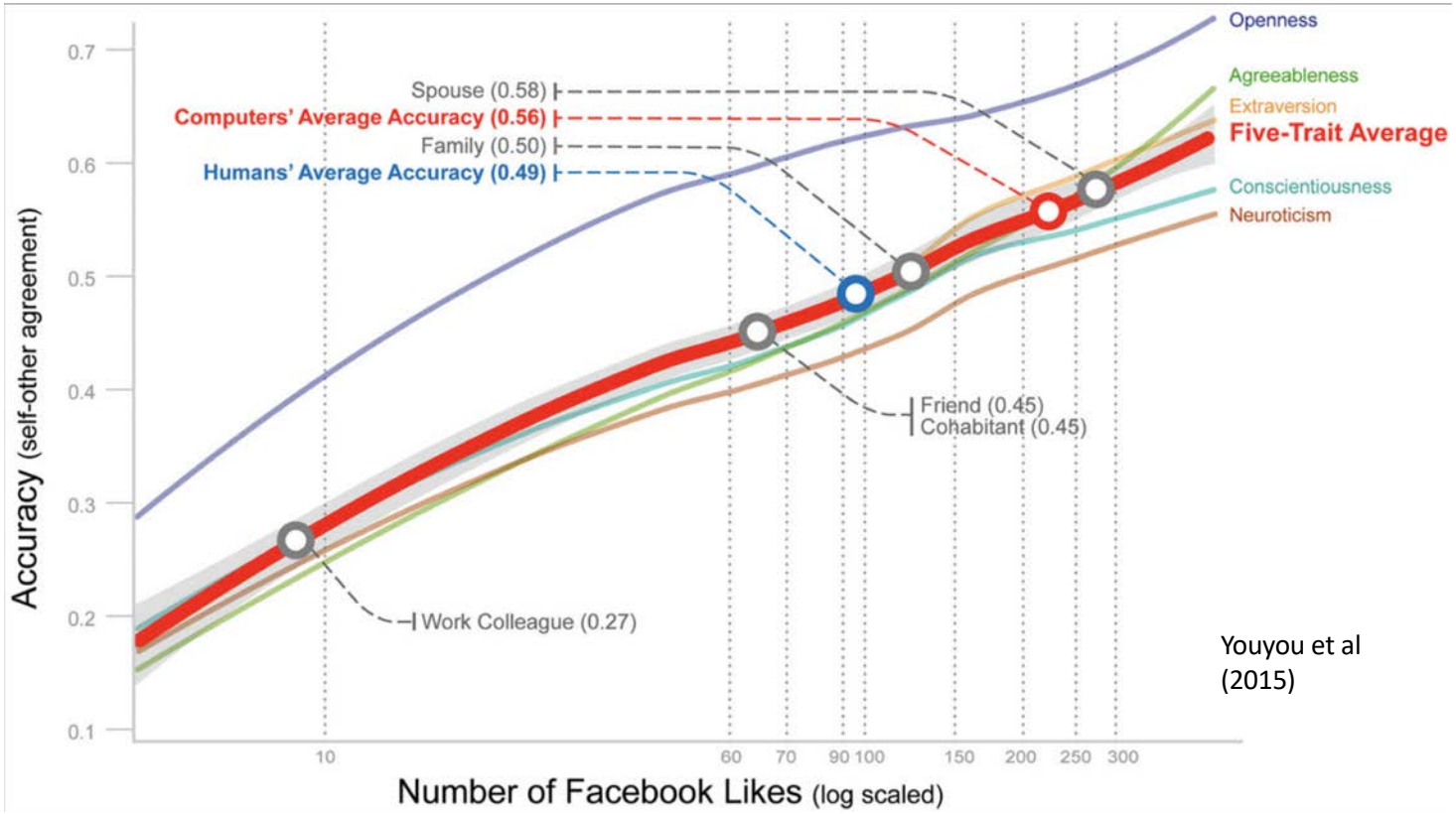
Data deposition: The data used in the study are shared with the academic community at mypersonality.org.

[1]W.Y. and M.K. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: yw541@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1418680112/-/DCSupplemental.

*The sample used in this study was obtained from the myPersonality project. myPersonality was a popular Facebook application that offered to its users psychometric tests and feedback on their scores. Since the data are secondary, anonymized, was previously published in the public domain, and was originally gathered with an explicit opt-in consent for reuse for research purposes beyond the original project, no IRB approval was needed. This was additionally confirmed by the Psychology Research Ethics Committee at the University of Cambridge.



Youyou et al (2015)

Sex orientation from photos

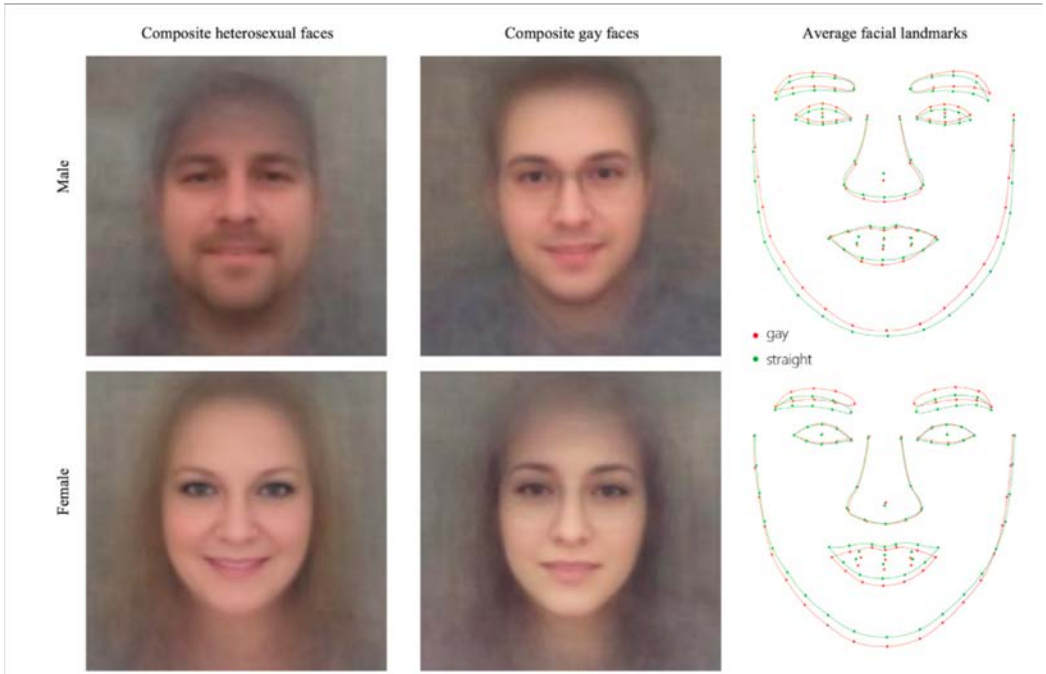


Figure 1. Composite faces and average face outlines produced by averaging faces/outlines classified as most likely to be gay or straight.

Michal Kosinski and Yilun Wang, *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images* (2017)

From procedure to substance

- Best data & tools are now in the private sector
  - How to address winner take all?
- Shoshana Zuboff: advances in ML are shifting the "division of learning" in society.
  - Who knows?
  - Who decides?
  - Who decides who decides?
    - THE AGE OF SURVEILLANCE CAPITALISM (2018)
- Cf. "freedom to observe," "freedom to learn"
  - Private decision space

Roadmap

- Level setting
- AI and GDPR
- **AI and military applications**
- AI and quantum computing

Strategic landscape

- Competitive adoption of autonomy, speed of conflict quickening
- Degrees of “autonomy” & automation
  - Decision to attack
  - Target selection
  - “loitering”
- High level norm erosion--blurring lines of “war” and “peace”
- Different strategic posture in Europe
  - Presence of electronic warfare/GPS denial
  - Triggers crisis in communications, chain of command

Investigation Report

Formal Investigation into the
Circumstances Surrounding the
Downing of Iran Air Flight 655
on 3 July 1988

Conflict in the information domain

- Historical trend of increasing individuality
- Increasingly able to shape, even invent, our own reality

Literature largely ignores military applications, lessons

- ”Autonomy” in systems back to the 1980s
  - Draw in carefully-studied lessons from accidents
- Cybersecurity
  - Defense & offense
- Battlefield cyber (look at the Snowden documents carefully)
  - Radio-delivered payloads
  - Attacks on other devices

Roadmap

- Level setting
- AI and GDPR
- AI and military applications
- **AI and quantum computing**

Second quantum revolution

- Current research agenda
- Advances in metrology, communications, sensing and computing
- Strategic race—EU has invested 1b

Simulating Physics with Computers

Richard P. Feynman

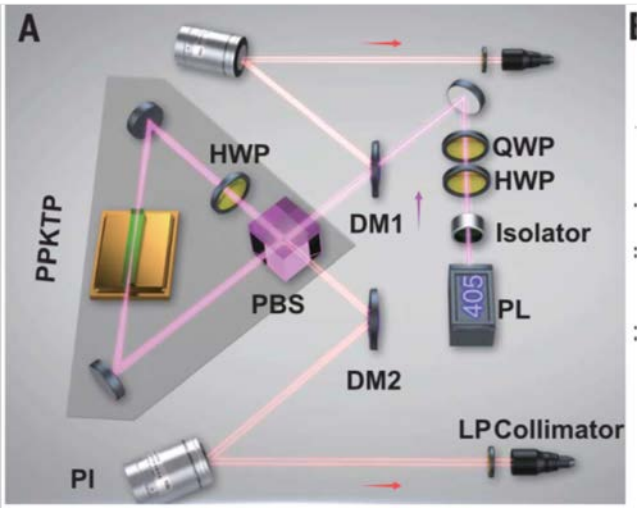
Department of Physics, California Institute of Technology, Pasadena, California 91107

Received May 7, 1981

1. INTRODUCTION

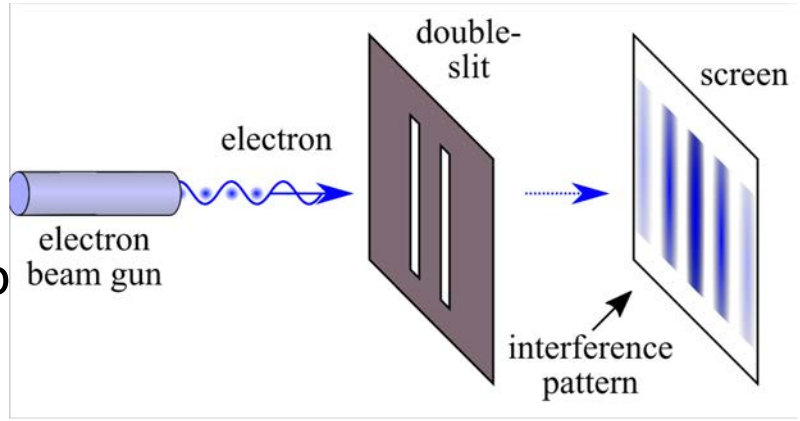
QUANTUM OPTICS

Satellite-based entanglement distribution over 1200 kilometers



Quantum background

- Characteristics of the subatomic world; human experience rarely encounters its strange physics
- Superposition: enables a form of parallel computing, *quantum parallelism*
- No-cloning theorem: can be leveraged to detect eavesdropping
- Entanglement: enables sensing at a distance, teleportation



Affordances

- Need for supercooling
- Cloud-based infrastructures (IBM, DWave)
- Expensive
- Decoherence
- Computing still hasn’t reached “quantum superiority”
  - Mass decryption may still be decades off

Implications for ML

- New sensing abilities (including at a distance)
  - MRI
  - Two-photon applications
- Search (Grover)
- Optimization (DWave)
- Simulation (IBM)
- AI (Google)
- Debugging in reverse because of quantum parallelism

Roadmap

- Level setting
- AI and GDPR
- AI and military applications
- AI and quantum computing
- **Thank you** ☺
- Chris Hoofnagle, choofnagle@berkeley.edu