Some implications of WTO ecommerce proposals restricting access to algorithms on algorithmic transparency

Dr. Ansgar Koene, University of Nottingham, UK [ansgar.koene@nottingham.ac.uk]

Some ecommerce proposals at the World Trade Organization would restrict the ability of regulators and experts to check algorithms (and source code) for bias or discrimination.[1] This note outlines some of the reasons why algorithmic transparency is important.

Algorithmic systems are increasingly at the heart of the digital economy, transforming diverse data sets into actionable recommendations; providing increasing levels of autonomy to cyber-physical systems, such as autonomous vehicles and the Internet of Things; and enabling tailor-made solutions for anything from healthcare to insurance and public services. At the same time, there is growing evidence that, opaque complex algorithmic systems can exhibit unintended and/or unjustified biases or errors with potentially significant consequences. The likelihood of such undesired outcomes is greatly increased when systems are deployed under novel operating conditions, such as in new environments or social-cultural contexts.

Algorithms "are inescapably value-laden. Operational parameters are specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others" [Mittelstadt et al. 2016]. Human values are (often unconsciously) embedded into algorithms during the process of design through the decisions of what categories and data to include and exclude. These values are highly subjective – what can appear 'neutral' or 'rational' to one person can seem unfair or discriminatory to another.

Due to the strongly interconnected and integrated nature of technical systems employed in the digital economy, clear accountability for bias and errors in products and services will require increased levels of auditability and transparency, which currently are often lacking.

When linked with pervasive and automated data collection (e.g. Internet of Things), where people implicitly provide the data that is used by the algorithmic system simply by being in the presence of the device, it can become difficult or impossible for individuals to identify which data were used to reach particular decision outcomes, and thus impossible to correct faulty data or assumptions.

Accordingly, there is now a growing demand for fairness, accountability, and transparency from algorithmic systems, and a growing research community (e.g. FAT* [www.fatml.org]) which is investigating how to deliver answers to these demands. When considering algorithmic fairness, it is important to remember that potential bias in training/validation data sets isn't the only source of

[1] See for example JOB/GC/178 and JOB/GC/177 from https://docs.wto.org/dol2fe/Pages/FE_Search/FE_S_S001.aspx

Arkansas algorithmic Medicaid assessment instrument

Upon introduced of the algorithmic assessment instrument in 2016, many people with cerebral palsy had their care dramatically reduced – they sued the state resulting in a court case.

Detailed examination of the assessment algorithm and data in court revealed:

The algorithm relied on 60 answer scores to questions about descriptions, symptoms and ailments. A small number of variables could matter enormously: a difference between scoring a three instead of a four on a handful of items meant a cut of dozens of care hours a month.

One variable was “foot problems”. Some assessors wrote that the person being assessed didn’t have any “foot problems” — when they were amputees and didn’t have feet.

The third-party software vendor implementing the system was found to have mistakenly used a version of the algorithm that didn’t account for diabetes issues.

Cerebral palsy, wasn’t properly coded in the algorithm, causing incorrect calculations for hundreds of people, mostly lowering their hours.

https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

possible bias. It can also be introduced through inappropriate data handling, inappropriate model selection, or incorrect algorithm design. Bias can also affect usage data.

Algorithmic systems should therefore be transparent to scrutiny whenever they play a role in any situation where a human would be legally required to provide an explanation for their decision. This approach prevents otherwise legally accountable decision-makers from “hiding” behind Algorithmic Decision Systems. While imperfect in its implementation, this was the intent behind Article 22 of the GDPR (Right to an explanation).

Depending on which aspect of an algorithmic system is in question, the meaning of “transparency” can be different:

1. The transparency of the systems’ *algorithms* can refer to a third party code review, analysis of how the algoriths works, inspection of internal and external bug reports, or assurance the software development processes are sound.
2. The transparency of the *data* used by the algorithmic system -- in particular by machine learning and deep learning algorithms -- can refer to the raw data, to the data’s sources, to how the data were preprocessed, to the methods by which it was verified as unbiased and representative (including looking for features that are proxies for information about protected classes, like race, that legally prohibited from being used), or to the processes by which the data are updated and the system recalibrated on them.
3. Algorithmic systems can also be transparent about their *goals*. When a system has multiple goals, this would mean being transparent about their relative priorities. For example, the artificial intelligence (AI) driving autonomous vehicles (AVs) might be aimed at reducing traffic fatalities, lowering the AVs’ environmental impact, reducing serious injuries, shortening transit times, and/or

avoiding property damage. A manufacturer could be required to be transparent about those goals and their priority.

4. Manufacturers or operators could be required to be transparent about the *outcomes* of the deployment of their algorithmic systems, including the internal states of the system (how worn are the brakes of an AV? how much electricity used?), the effects on external systems (how many accidents, or times it has caused another AV to swerve?), and computer-based interactions with other algorithmic systems (what communications with other AVs, what data fed into traffic monitoring systems?).
5. Manufacturers or operators may be required to be transparent about their overall *compliance* with whatever transparency requirements have been imposed upon them. In many instances, there may be a requirement that these compliance reports are backed by data that is inspectable by regulators or the general public.

Note that "transparency" has different meanings in this categorization. It can mean: access upon request to the public or authorized people; public posting of information; direct inspection of internal processes; delivery of complete subsystems and their data for testing by authorized people, with the results reported to the public or to regulatory bodies; access to computer scientists and managers to explain algorithmic or operational processes.

* * *

Algorithmic systems for decision making require clear mechanisms of accountability due to their potential to bring about consequences that are detrimental on a number of levels:

- *detrimental to the individual:* individual citizens might become the recipients of inaccurate decisions or be treated more harshly in comparison to others. Where this relates to decisions over, for instance, prison sentences[2], this can have very serious consequences. Individuals might also receive false/misleading/skewed information e.g. as a result of online searches and this can alter their perceptions or behaviours, perhaps including their voting behaviours[3]. The collection and collation of information necessitated by some algorithmic processes might also be considered a breach of privacy.
- *detrimental to groups:* where algorithmic processes appear to produce different results for different (demographic) groups, this often places some of those groups at a disadvantage. For instance, the case studies below suggest that blacks might be more vulnerable than whites to longer prison sentences, lack of access to facial recognition technologies, stereotyping in online advertisements, and stereotyped/prejudicial representations in online searches. This can have further detrimental consequences for those groups if the outcomes of those processes reinforce wider societal prejudices.
- *detrimental to society:* entire societies are disadvantaged if the outcomes of algorithmic processes cannot be relied on to be accurate and/or neutral. Incorrect decisions can have societal effects –

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] https://www.theguardian.com/technology/2017/jul/31/facebook-dark-ads-can-swing-opinions-politics-research-shows

for instance the wrongful arrest of individuals based on facial recognition technologies places a society at risk if actual offenders are overlooked, and stereotyped online content risks reinforcing prejudices. Furthermore, these outcomes may lead to loss of trust amongst the population as well as concerns that companies utilising these systems are allowed too much power.

In order to guard against these potential detrimental consequences, it is important to be able to inspect an algorithmic system's *data* and *algorithms* to:

- Check for bias in the data and algorithms that affects the fairness of the system.
- Check that the system is drawing inferences from relevant and representative data.
- See if we can learn anything from the machine's way of connecting and weighting the data -- perhaps there's a meaningful correlation we had not been aware of.
- Look for, and fix, bugs.
- Guard against malicious/adversarial data injection[4].

This requires the hierarchy of *goals* and *outcomes* to be transparent so:

- They can be debated and possibly regulated.
- Regulators and the public can assess how well an algorithmic system has performed relative to its goals and compared to the pre-algorithmic systems it may be replacing or supplementing.

Governance of Algorithmic Decision-Making systems

The development of governance frameworks for Algorithmic Decision Making is still in its infancy. Both the development of industry standards and government regulations have not yet matured to a level that can provide clarity about the kind of algorithm transparency that will be necessary to satisfy future product/service quality assurance requirements.

International industry standards development

In 2017, the Institute of Electrical and Electronics Engineers (IEEE.the world's largest technical professional association) was the first of the international standards setting bodies to launch a programme for developing standards specifically related to the ethics and social impact of algorithmic decision making. As part of the IEEE Global Initiative for Ethics of Autonomous and Intelligent Systems, the P7000-series of standards was initiated which currently includes 13 standards development working groups. The standards that are currently in development include:

- IEEE P7000: Model Process for Addressing Ethical Concerns During System Design
- IEEE P7001: Transparency of Autonomous Systems
- IEEE P7003: Algorithmic Bias Considerations
- IEEE P7009: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

The earliest of these is expected to reach completion in the second half of 2019.

[4] https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

At the start of 2018, ISO/IEC initiated the ISO/IEC JTC/1 SC42 subcommittee to develop standards related to Artificial Intelligence. This standards development effort is currently still at the stage of study groups that are investigating the need and feasibility of developing standards for specific AI related issues (e.g. trustworthiness). Completed ISO/IEC JTC/1 SC42 standards ae unlikely to appear before 2022.

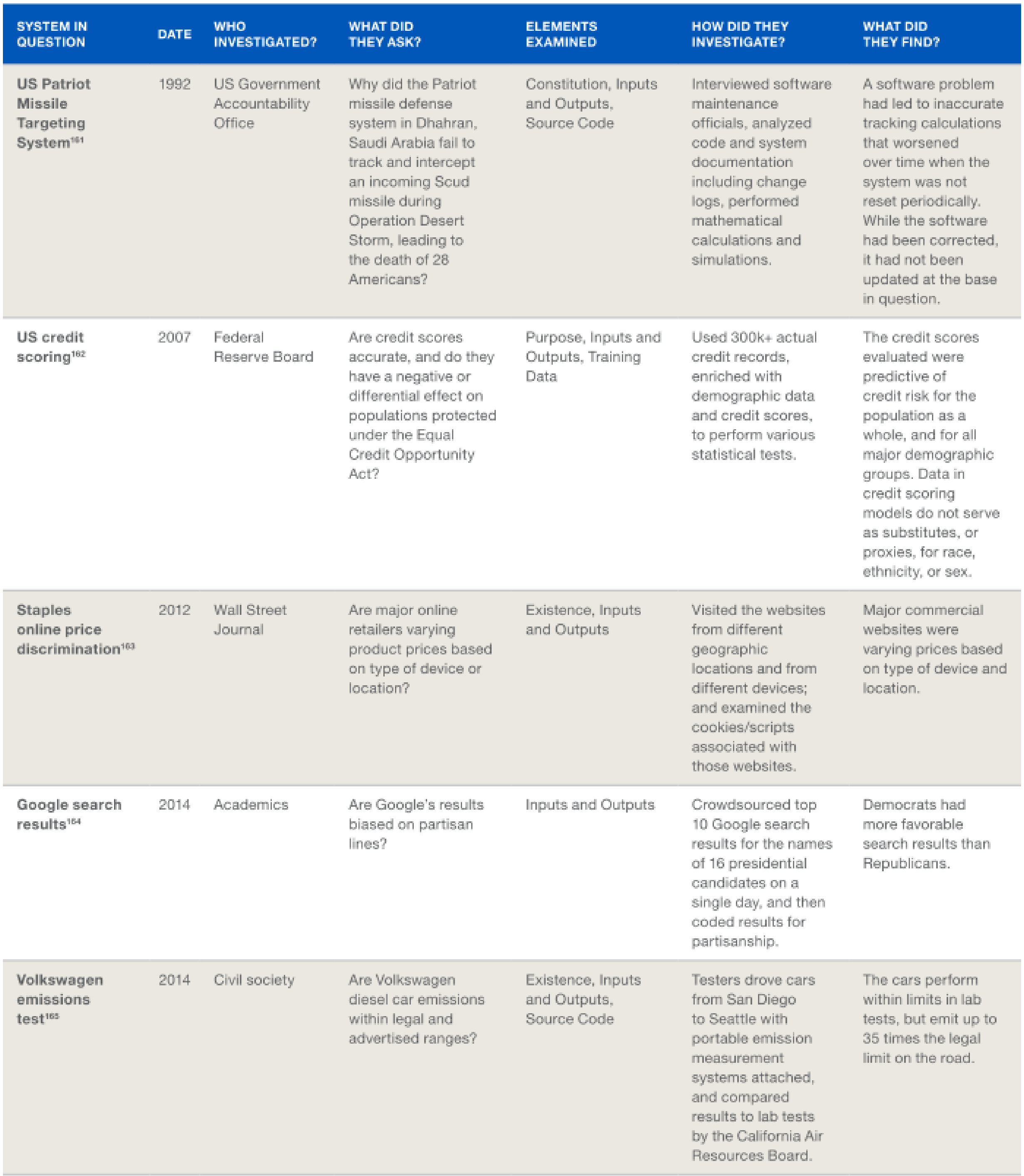
Government regulation

Most national governments, as well as the European Commission, are still engaged in exploratory inquires to try to understand what kind of legislation might be required in order to protect their citizens against detrimental consequences of bad algorithmic decision-making. For example:

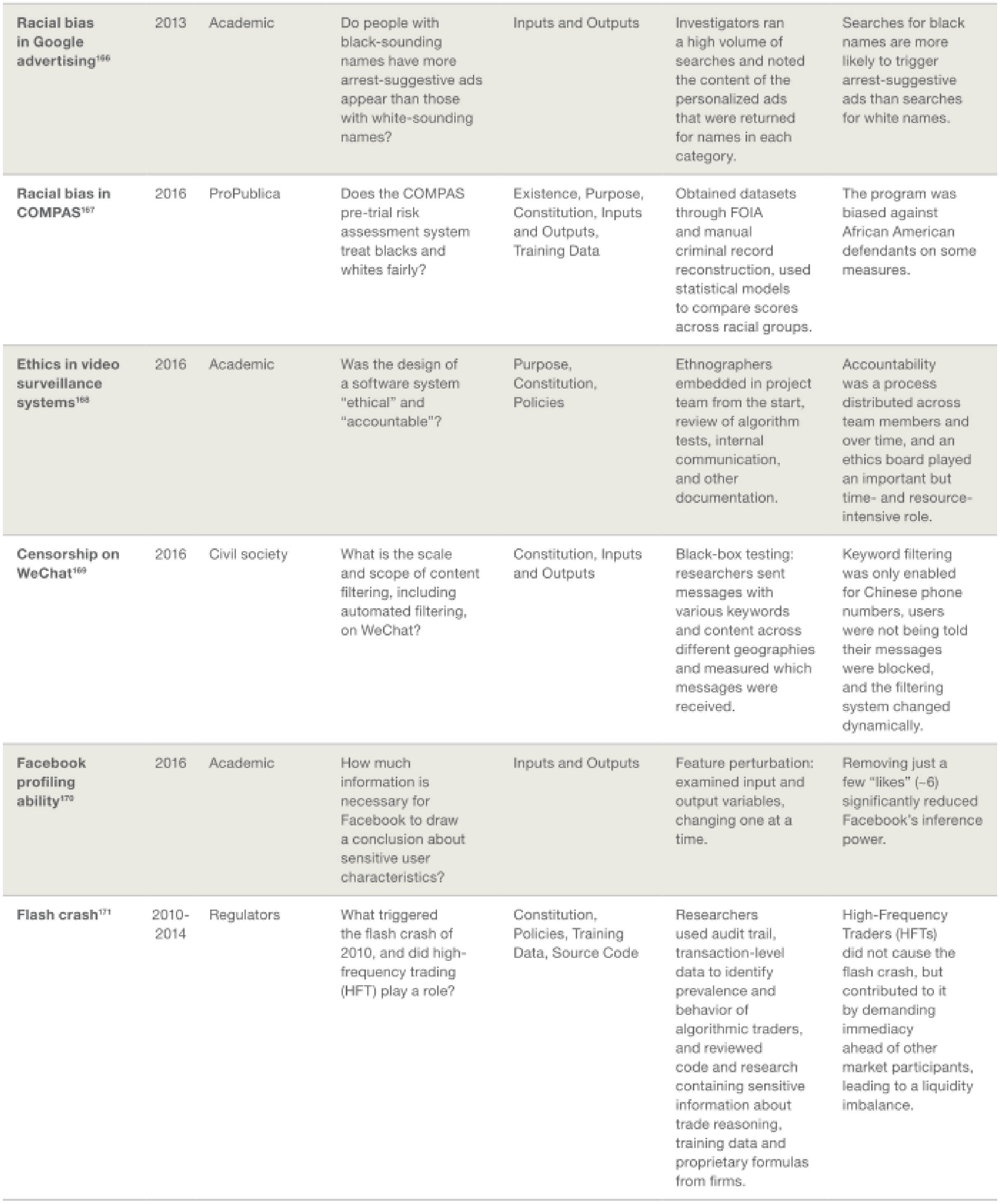
- In the UK, the a new government Center for Data Ethics and Innovation has been establish to lead policy development on AI. The public consultation seeking views on its work and activities closed on 5 September 2018.
- On 14 June 2018, the European Commission established a High-Level Expert group on Artificial Intelligence, supported by a European AI Alliance, to help the European Commission implement its European strategy on AI, which aims to establish "AI ethics guidelines" and "Guidance on the interpretation of the Product Liability directive" in 2019.
- On June 5th 2018, the Personal Data Protection Commission of Singapore published a "Discussion paper on AI and Personal Data – Fostering Responsible Development and Adoption of AI" as a first step towards establishing its regulatory framework for AI.

Examples of public scrutiny of automated decisions

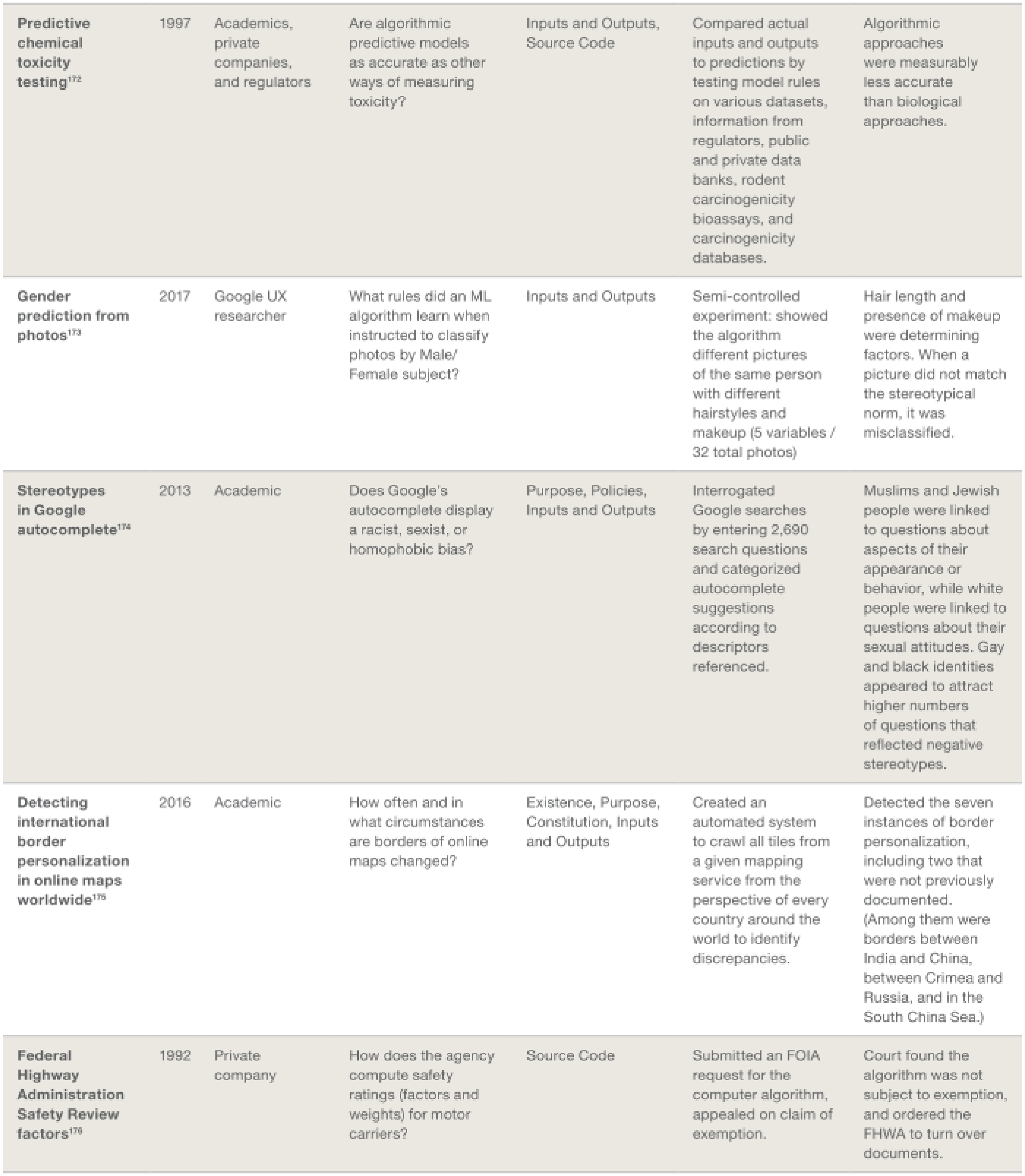
[https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods]



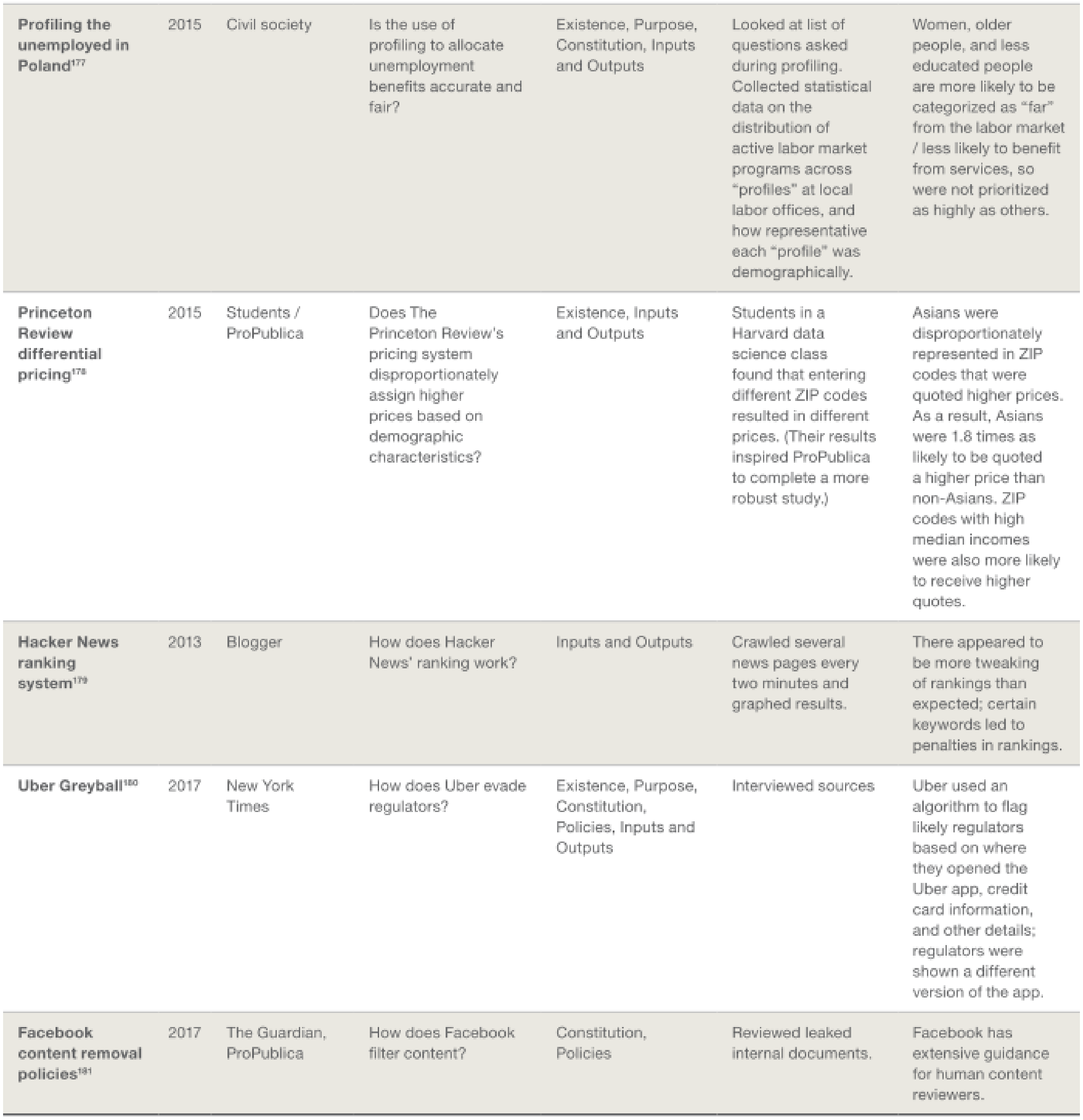
SYSTEM IN QUESTION	DATE	WHO INVESTIGATED?	WHAT DID THEY ASK?	ELEMENTS EXAMINED	HOW DID THEY INVESTIGATE?	WHAT DID THEY FIND?
US Patriot Missile Targeting System[161]	1992	US Government Accountability Office	Why did the Patriot missile defense system in Dhahran, Saudi Arabia fail to track and intercept an incoming Scud missile during Operation Desert Storm, leading to the death of 28 Americans?	Constitution, Inputs and Outputs, Source Code	Interviewed software maintenance officials, analyzed code and system documentation including change logs, performed mathematical calculations and simulations.	A software problem had led to inaccurate tracking calculations that worsened over time when the system was not reset periodically. While the software had been corrected, it had not been updated at the base in question.
US credit scoring[162]	2007	Federal Reserve Board	Are credit scores accurate, and do they have a negative or differential effect on populations protected under the Equal Credit Opportunity Act?	Purpose, Inputs and Outputs, Training Data	Used 300k+ actual credit records, enriched with demographic data and credit scores, to perform various statistical tests.	The credit scores evaluated were predictive of credit risk for the population as a whole, and for all major demographic groups. Data in credit scoring models do not serve as substitutes, or proxies, for race, ethnicity, or sex.
Staples online price discrimination[163]	2012	Wall Street Journal	Are major online retailers varying product prices based on type of device or location?	Existence, Inputs and Outputs	Visited the websites from different geographic locations and from different devices; and examined the cookies/scripts associated with those websites.	Major commercial websites were varying prices based on type of device and location.
Google search results[164]	2014	Academics	Are Google's results biased on partisan lines?	Inputs and Outputs	Crowdsourced top 10 Google search results for the names of 16 presidential candidates on a single day, and then coded results for partisanship.	Democrats had more favorable search results than Republicans.
Volkswagen emissions test[165]	2014	Civil society	Are Volkswagen diesel car emissions within legal and advertised ranges?	Existence, Inputs and Outputs, Source Code	Testers drove cars from San Diego to Seattle with portable emission measurement systems attached, and compared results to lab tests by the California Air Resources Board.	The cars perform within limits in lab tests, but emit up to 35 times the legal limit on the road.



Racial bias in Google advertising[166]	2013	Academic	Do people with black-sounding names have more arrest-suggestive ads appear than those with white-sounding names?	Inputs and Outputs	Investigators ran a high volume of searches and noted the content of the personalized ads that were returned for names in each category.	Searches for black names are more likely to trigger arrest-suggestive ads than searches for white names.
Racial bias in COMPAS[167]	2016	ProPublica	Does the COMPAS pre-trial risk assessment system treat blacks and whites fairly?	Existence, Purpose, Constitution, Inputs and Outputs, Training Data	Obtained datasets through FOIA and manual criminal record reconstruction, used statistical models to compare scores across racial groups.	The program was biased against African American defendants on some measures.
Ethics in video surveillance systems[168]	2016	Academic	Was the design of a software system "ethical" and "accountable"?	Purpose, Constitution, Policies	Ethnographers embedded in project team from the start, review of algorithm tests, internal communication, and other documentation.	Accountability was a process distributed across team members and over time, and an ethics board played an important but time- and resource-intensive role.
Censorship on WeChat[169]	2016	Civil society	What is the scale and scope of content filtering, including automated filtering, on WeChat?	Constitution, Inputs and Outputs	Black-box testing: researchers sent messages with various keywords and content across different geographies and measured which messages were received.	Keyword filtering was only enabled for Chinese phone numbers, users were not being told their messages were blocked, and the filtering system changed dynamically.
Facebook profiling ability[170]	2016	Academic	How much information is necessary for Facebook to draw a conclusion about sensitive user characteristics?	Inputs and Outputs	Feature perturbation: examined input and output variables, changing one at a time.	Removing just a few "likes" (~6) significantly reduced Facebook's inference power.
Flash crash[171]	2010-2014	Regulators	What triggered the flash crash of 2010, and did high-frequency trading (HFT) play a role?	Constitution, Policies, Training Data, Source Code	Researchers used audit trail, transaction-level data to identify prevalence and behavior of algorithmic traders, and reviewed code and research containing sensitive information about trade reasoning, training data and proprietary formulas from firms.	High-Frequency Traders (HFTs) did not cause the flash crash, but contributed to it by demanding immediacy ahead of other market participants, leading to a liquidity imbalance.



Predictive chemical toxicity testing[172]	1997	Academics, private companies, and regulators	Are algorithmic predictive models as accurate as other ways of measuring toxicity?	Inputs and Outputs, Source Code	Compared actual inputs and outputs to predictions by testing model rules on various datasets, information from regulators, public and private data banks, rodent carcinogenicity bioassays, and carcinogenicity databases.	Algorithmic approaches were measurably less accurate than biological approaches.
Gender prediction from photos[173]	2017	Google UX researcher	What rules did an ML algorithm learn when instructed to classify photos by Male/Female subject?	Inputs and Outputs	Semi-controlled experiment: showed the algorithm different pictures of the same person with different hairstyles and makeup (5 variables / 32 total photos)	Hair length and presence of makeup were determining factors. When a picture did not match the stereotypical norm, it was misclassified.
Stereotypes in Google autocomplete[174]	2013	Academic	Does Google's autocomplete display a racist, sexist, or homophobic bias?	Purpose, Policies, Inputs and Outputs	Interrogated Google searches by entering 2,690 search questions and categorized autocomplete suggestions according to descriptors referenced.	Muslims and Jewish people were linked to questions about aspects of their appearance or behavior, while white people were linked to questions about their sexual attitudes. Gay and black identities appeared to attract higher numbers of questions that reflected negative stereotypes.
Detecting international border personalization in online maps worldwide[175]	2016	Academic	How often and in what circumstances are borders of online maps changed?	Existence, Purpose, Constitution, Inputs and Outputs	Created an automated system to crawl all tiles from a given mapping service from the perspective of every country around the world to identify discrepancies.	Detected the seven instances of border personalization, including two that were not previously documented. (Among them were borders between India and China, between Crimea and Russia, and in the South China Sea.)
Federal Highway Administration Safety Review factors[176]	1992	Private company	How does the agency compute safety ratings (factors and weights) for motor carriers?	Source Code	Submitted an FOIA request for the computer algorithm, appealed on claim of exemption.	Court found the algorithm was not subject to exemption, and ordered the FHWA to turn over documents.



Profiling the unemployed in Poland[177]	2015	Civil society	Is the use of profiling to allocate unemployment benefits accurate and fair?	Existence, Purpose, Constitution, Inputs and Outputs	Looked at list of questions asked during profiling. Collected statistical data on the distribution of active labor market programs across "profiles" at local labor offices, and how representative each "profile" was demographically.	Women, older people, and less educated people are more likely to be categorized as "far" from the labor market / less likely to benefit from services, so were not prioritized as highly as others.
Princeton Review differential pricing[178]	2015	Students / ProPublica	Does The Princeton Review's pricing system disproportionately assign higher prices based on demographic characteristics?	Existence, Inputs and Outputs	Students in a Harvard data science class found that entering different ZIP codes resulted in different prices. (Their results inspired ProPublica to complete a more robust study.)	Asians were disproportionately represented in ZIP codes that were quoted higher prices. As a result, Asians were 1.8 times as likely to be quoted a higher price than non-Asians. ZIP codes with high median incomes were also more likely to receive higher quotes.
Hacker News ranking system[179]	2013	Blogger	How does Hacker News' ranking work?	Inputs and Outputs	Crawled several news pages every two minutes and graphed results.	There appeared to be more tweaking of rankings than expected; certain keywords led to penalties in rankings.
Uber Greyball[180]	2017	New York Times	How does Uber evade regulators?	Existence, Purpose, Constitution, Policies, Inputs and Outputs	Interviewed sources	Uber used an algorithm to flag likely regulators based on where they opened the Uber app, credit card information, and other details; regulators were shown a different version of the app.
Facebook content removal policies[181]	2017	The Guardian, ProPublica	How does Facebook filter content?	Constitution, Policies	Reviewed leaked internal documents.	Facebook has extensive guidance for human content reviewers.