



The State Bar of California

CLOSING

II.D. Presentation
08-11-21 CTJG Meeting
Open Session

WORKING GROUP

DATE: August 10, 2021

TO: Closing the Justice Gap Working Group

FROM: State Bar Staff

SUBJECT: II.D. Presentation: Technology Driven Legal Services Delivery Systems Including the Issues of: Dark Patterns; the Need for Scale in Algorithmic Systems; and the Current Limits of Such Systems (Mihir Kshirsagar – Princeton University Center for Information Technology Policy)

For Closing the Justice Gap Working Group August 11, 2021 meeting agenda item II.D. – *Presentation: Technology Driven Legal Services Delivery Systems Including the Issues of: Dark Patterns; the Need for Scale in Algorithmic Systems; and the Current Limits of Such Systems*, Mihir Kshirsagar, the guest speaker, has shared the following background materials for your review:

- Article entitled "[Dark Patterns: Past, Present, and Future](#)," Arvind Narayanan, Arunesh Mathur, Marshini Chetty, Mihir Kshirsagar, Communications of the Association for Computing Machinery (ACM), September, 2020;
- Article entitled "[Shining a Light on Dark Patterns](#)," Jamie Luguri, Lior Jacob Strahilevitz, Journal of Legal Analysis, Harvard Law School, March 23, 2021;
- Website post entitled "[Machine learning sucks at covid](#)," Pluralistic: Daily Links from Cory Doctorow, August 2, 2021; and
- Presentation slides for "How to recognize AI snake oil," Arvind Narayanan, Associate Professor of Computer Science, Princeton University Center for Information Technology Policy – attached.

How to recognize AI snake oil

Arvind Narayanan

Associate Professor of Computer Science

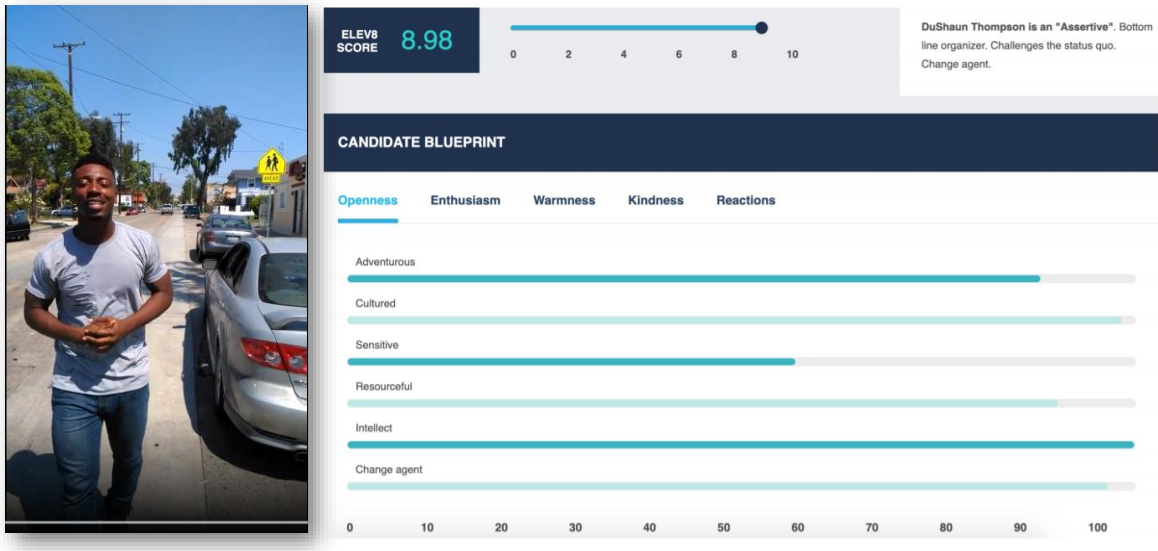
@random_walker



Much of what's being sold as "AI" today is snake oil — it does not and cannot work.

Why is this happening? How can we recognize flawed AI claims and push back?

Assessing personality & job suitability from 30-second video



Millions of people applying for jobs have been subjected to these types of algorithmic assessment systems. This is an actual video and screenshot from promo material of actual company. These systems claim to work by analyzing not even what the candidate said, but rather body language, speech patterns, etc.

Common sense tells you this isn't possible, and AI experts would agree. This product is essentially an elaborate random number generator.

<https://arxiv.org/abs/1906.09208>

Vendor name	Funding	# of employees	Location
8 and Above	–	1-10	WA, USA
ActiView	\$6.5M	11-50	Israel
Applied	£2M	11-50	UK
Assessment Innovation	\$1.3M	1-10	NY, USA
Good&Co	\$10.3M	51-100	CA, USA
Harver	\$14M	51-100	NY, USA
HireVue	\$93M	251-500	UT, USA
impress.ai	\$1.4M	11-50	Singapore
Knockri	–	11-50	Canada
Koru	\$15.6M	11-50	WA, USA
LaunchPad Recruits	£2M	11-50	UK
myInterview	\$1.4M	1-10	Australia
Plum.io	\$1.9M	11-50	Canada
PredictiveHire	A\$4.3M	11-50	Australia
pymetrics	\$56.6M	51-100	NY, USA
Scoutible	\$6.5M	1-10	CA, USA
Teamscope	€800K	1-10	Estonia
ThriveMap	£781K	1-10	UK
Yobs	\$1M	11-50	CA, USA

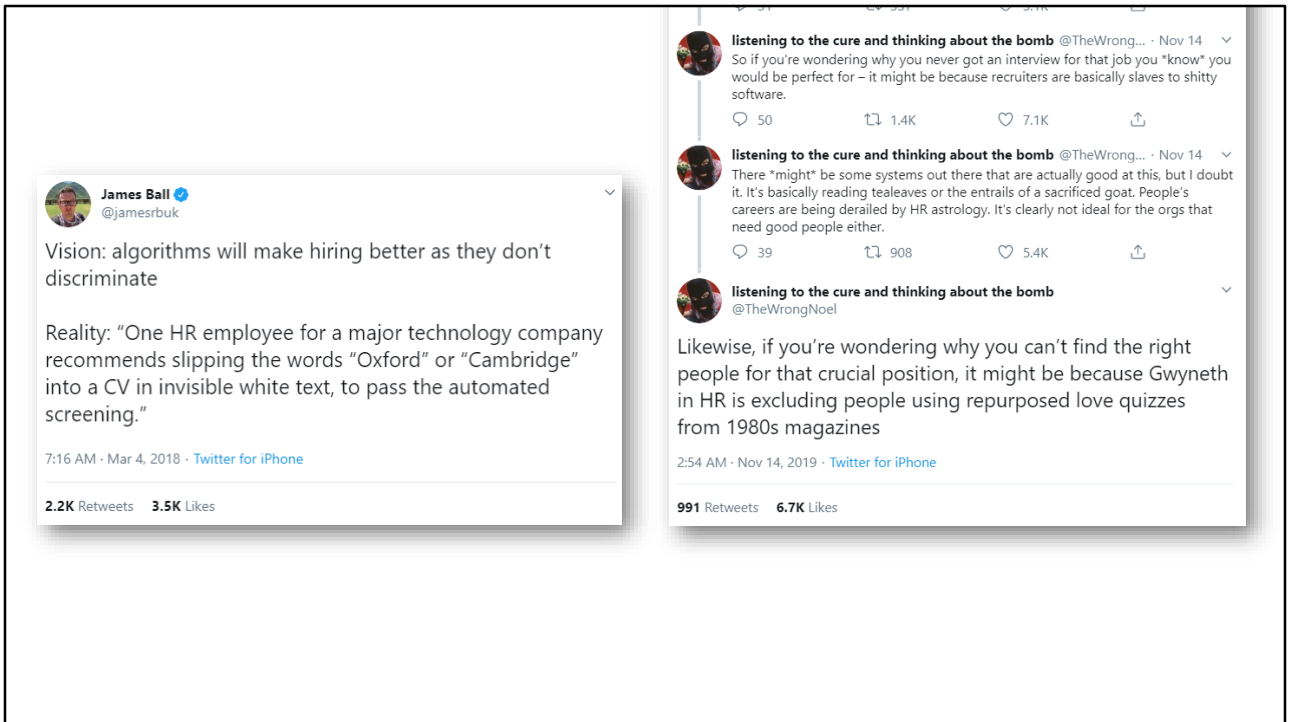
Not all these companies offer AI assessments of job candidates, but most do.

Raghavan et al. *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*. Preprint 2019.

These companies have collectively raised hundreds of millions of dollars and are going after clients aggressively. The phenomenon of job candidates being screened out by bogus AI is about to get much, much worse.

My goal today in putting up this slide is not to name names but show how widespread the problem is.

The table is from this excellent draft paper by Cornell researchers:
<https://arxiv.org/abs/1906.09208>



People are learning to work around these bogus systems and sharing horror stories on social media.

<https://twitter.com/jamesrbuk/status/970271658769571840>

<https://twitter.com/TheWrongNoel/status/1194842728862892033>

How did this happen? Why are HR departments apparently so gullible?

In what domains other than hiring are snake oil AI tools being sold?

Why is there so much AI snake oil?

AI is an umbrella term for a set of related technologies

Some of those technologies have made genuine, remarkable, widely-publicized progress

Companies exploit public confusion, slap the “AI” label on whatever they’re selling

This is my hypothesis for why there’s so much AI snake oil and why policy makers and decision makers are falling for it.

For example, AlphaGo is a remarkable intellectual accomplishment that deserves to be celebrated. Ten years ago, most experts would not have thought it possible.

But it has nothing in common with a tool that claims to predict job performance.

Massive effort to influence public opinion

Result:

6.1 The public predicts a 54% likelihood of high-level machine intelligence within 10 years

Respondents were asked to forecast when high-level machine intelligence will be developed. High-level machine intelligence was defined as the following:

We have high-level machine intelligence when machines are able to perform almost all tasks that are economically relevant today better than the median human (today) at each task. These tasks include asking subtle common-sense questions such as those that travel agents would ask. For the following questions, you should ignore tasks that are legally or culturally restricted to humans, such as serving on a jury.¹³

Respondents were asked to predict the probability that high-level machine intelligence will be built in 10, 20, and 50 years.

Zhang & Dafoe. *Artificial Intelligence: American Attitudes and Trends*. 2019.

Companies advertising AI as the solution to all problems have been helped along by credulous media. As a result, the US public believes that the automation of all jobs is only 10 years away! https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf

If policy makers believe that such a radical transformation is imminent, imagine how it would warp our priorities. I believe that that's actually happening today.

Note that AI experts have a more modest estimate that Artificial General Intelligence or Strong AI is about 50 years away, but history tells us that even experts tend to be wildly optimistic about AI predictions.

Genuine, rapid technological progress

- Content identification (Shazam, reverse image search)
- Face recognition*
- Medical diagnosis from scans
- Speech to text
- Deepfakes*

Perception

* Ethical concerns because of high accuracy

Let's get more concrete. I'll break down AI applications into three (non-exhaustive) categories in this and the next few slides.

Everything on this slide is a perception problem (Deepfakes are not purely perception but closely related. They are created by having a generative neural network and a discriminative/perceptive neural network compete with each other). Perception is one of a few areas in which AI has made rapid progress.

AI is already at or beyond human accuracy in all the tasks on this slide and is continuing to get better rapidly.

The fundamental reason for progress is that there is no uncertainty or ambiguity in these tasks -- given two images of faces, there's ground truth about whether or not they represent the same person. So, given enough data and compute, AI will learn the patterns that distinguish one face from another. There have been some notable failures of face recognition, but I'm comfortable predicting that it will continue to get much more accurate (and that's exactly why we should worry).

Far from perfect, but improving

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

Automating
judgment

Ethical concerns in part because some error is inevitable

This second category is about applications that try to automate judgment. Humans have some heuristic in our minds, such as what is spam and not spam, and given enough examples, the machine tries to learn it.

AI will never be perfect at these tasks because they involve judgment and reasonable people can disagree about the correct decision.

We seem to have decided to adopt these systems and must decide how best to govern them, such as figuring out due process mechanisms for people whose content gets incorrectly taken down.

Fundamentally dubious

- Predicting criminal recidivism
- Predicting job performance
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Predicting
social outcomes

Ethical concerns amplified by inaccuracy

I will focus the rest of my talk on this third category, where there's a lot of snake oil.

I already showed you tools that claim to predict job suitability. Similarly, bail decisions are being made based on an algorithmic prediction of recidivism. People are being turned away at the border based on an algorithm that analyzed their social media posts and predicted a terrorist risk.

These problems are hard because we can't predict the future. That should be common sense. But we seem to have decided to suspend common sense when AI is involved.

Incomplete & crude but useful breakdown

Genuine, rapid progress

- Shazam, reverse img search
- Face recognition
- Med. diagnosis from scans
- Speech to text
- Deepfakes

Perception

Imperfect but improving

- Spam detection
- Copyright violation
- Automated essay grading
- Hate speech detection
- Content recommendation

Automating
judgment

Fundamentally dubious

- Predicting recidivism
- Predicting job success
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Predicting
social outcomes

This is of course not even close to an exhaustive list of things that AI is used for (not on this list: all of robotics, game playing, ...). The point, however, is to illustrate how limits to accuracy are quantitatively and qualitatively different for different types of tasks.

I'll show you that there has been no real improvement in the third category, despite how much data you throw at it.

Can social outcomes be predicted?



Matthew Salganik, Ian Lundberg, Alex Kindel, Sara McLanahan, et al.

Mass collaboration involving 457 researchers.

This is the most rigorous effort that I'm aware of to measure the predictability of social outcomes.

[Description of Fragile Families Study and Challenge.]

<https://fragilefamilies.princeton.edu/about>

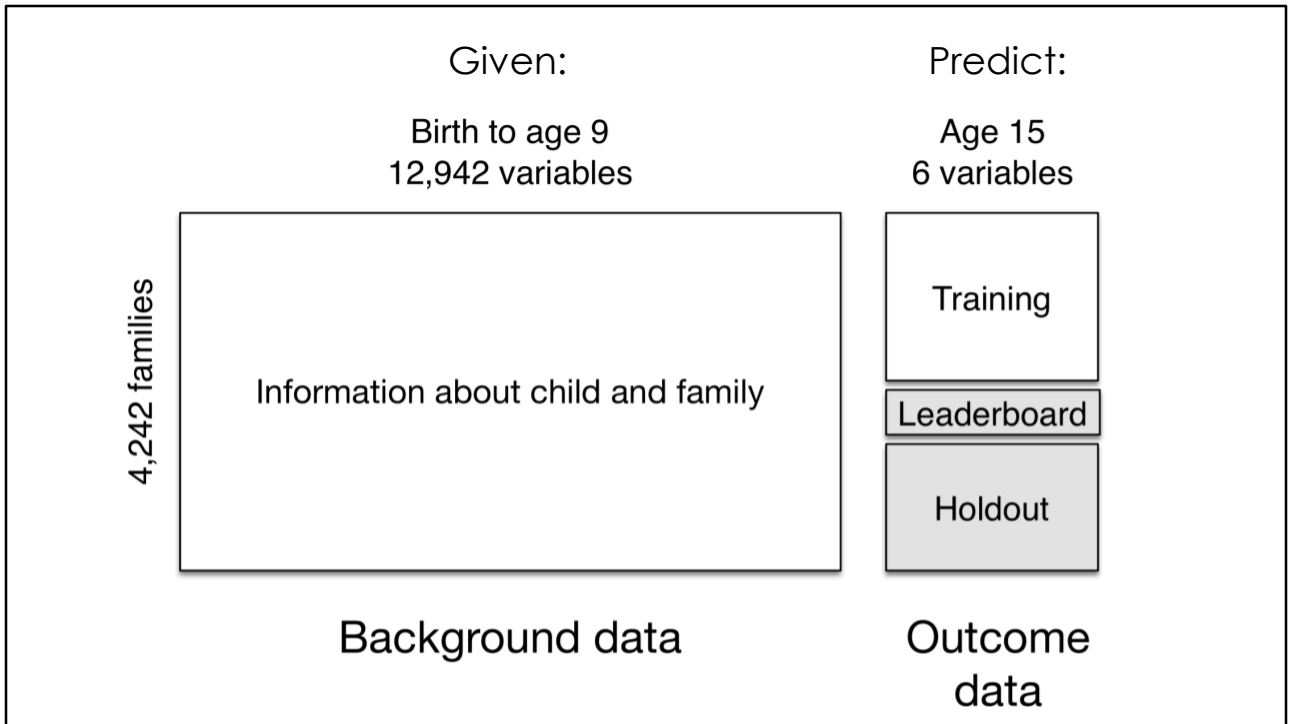
Birth to age 9
12,942 variables

4,242 families

Information about child and family

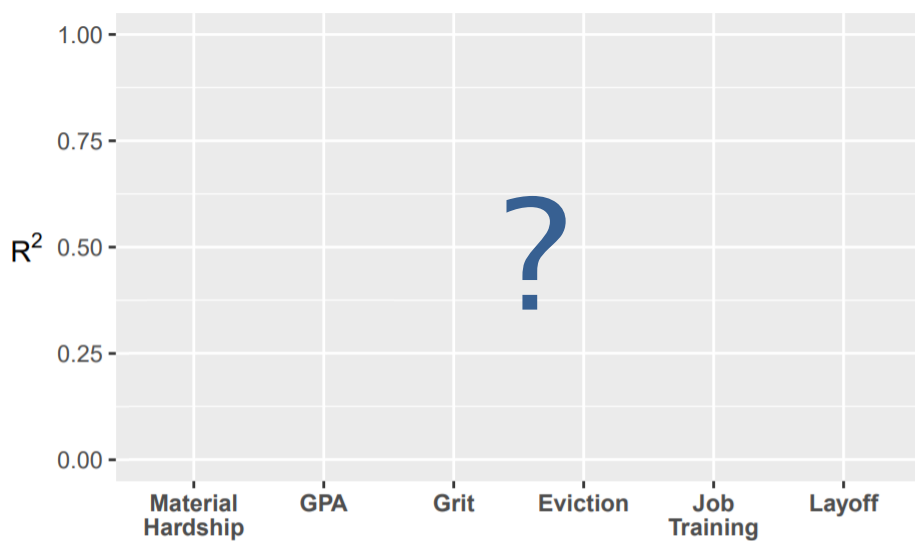
They collected a huge amount of data about each child and family based on in-depth interviews and in-home observation repeated several times over many years.

Slides from Matt Salganik used with permission.



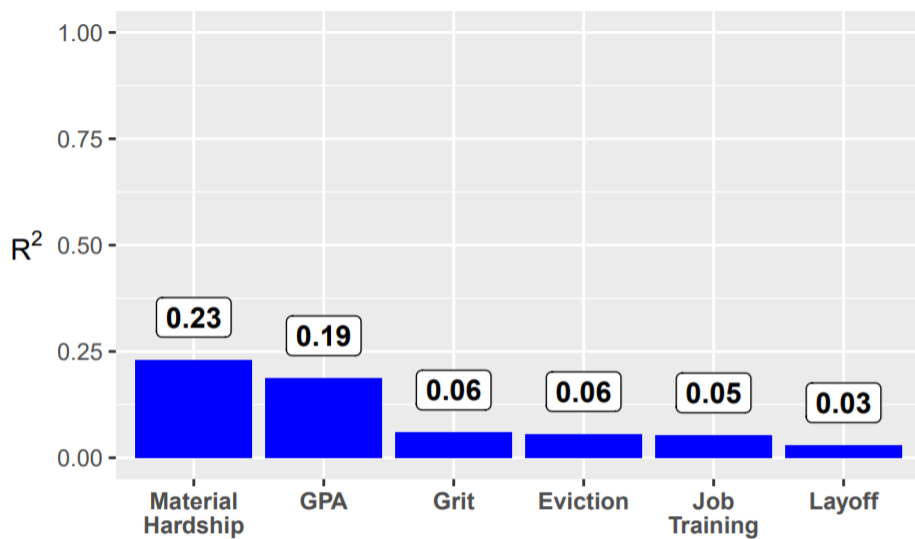
They used a challenge setting similar to many other machine learning competitions.

The task is to learn the relationship between the background data and the outcome data based on the training instances. Accuracy is assessed on the leaderboard during the competition and evaluated on the held-out data after the end of the competition.



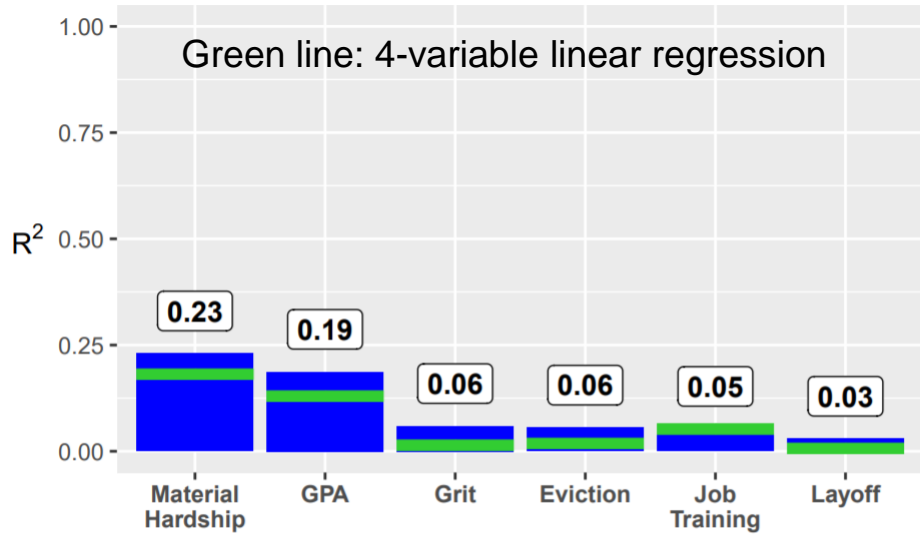
Perfect prediction corresponds to an R^2 of 1. Predicting the mean for every instance corresponds to an R^2 of 0 (i.e. the model has not at all learned to discriminate between instances).

Most people's intuition seems to suggest R^2 values between 0.5 and 0.8. Many of the experts organizing the challenge had high expectations.



This is what actually happened.

Remember: hundreds of trained AI/ML researchers and students attempted this, were incentivized to maximize predictive accuracy --- both because of the leaderboard and appeal to noble intentions --- and were given 13,000 features per family. These were the best performing models.



13,000 features hardly better than 4 features!
“AI” hardly better than simple linear formula

This is the crux.

Regression analysis is a hundred years old.

Accuracy of recidivism prediction

COMPAS tool (137 features): $65\% \pm 1\%$ (slightly better than random)
Logistic regression (2 features): $67\% \pm 2\%$



Age and number of priors

Dressel & Farid. *The accuracy, fairness, and limits of predicting recidivism*. Science Advances 2018.

The same sort of finding is repeated in many domains.

Note that this is accuracy and not R^2 , so 65% is only slightly better than random.

The actual accuracy is probably even lower, because while the tool claims to predict recidivism, it actually predicts re-arrest, because that's what is recorded in the data. So at least some of the predictive performance of the algorithm comes from being able to predict the biases of policing.

<https://advances.sciencemag.org/content/4/1/eaao5580>

Key claim

For predicting social outcomes, AI is not substantially better than manual scoring using just a few features

Related: Jung et al. *Simple Rules for Complex Decisions*. 2017.

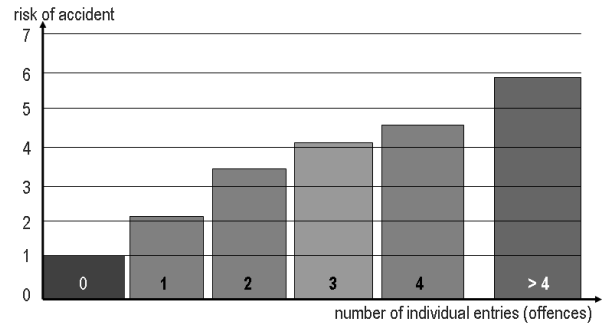
This is a falsifiable claim. Of course, I'm willing to change my mind or add appropriate caveats to the claim if contrary evidence comes to light.

But given the evidence so far, this seems the most prudent view.

This paper makes a related claim: <https://arxiv.org/abs/1702.04690>

This is already widely applied in some domains

Violations	Points
Failure to observe traffic signals (e.g., red light)	3
Speeding 1-9 mph above the posted speed limit	3
Illegal turn (e.g., U-turn)	3
Reckless driving	4
Speeding 10-19 mph above posted speed limit	4
Unsafe passing	4
Failure to yield right-of-way	4
Tailgating	4
Failure to obey railroad crossing signal	4
Driving Under the Influence (DUI)	6
Speeding 20 mph above speed limit	6
Speeding in excess of 80 mph	6



*) Deutschland 2005. Nach Schade, F.-D. (2006). The effectiveness of the Points System in Germany. Fik-to-drive-Tagungsband 2006.

Demerits on a driver's license can be seen as a way of predicting accident risk. Some studies have found such systems to be reasonably well calibrated.

[This didn't make it into the talk] We've known for a long time that in many domains, if all we really want to do is prediction (often it's not), then simple formulas are more accurate than predictions by humans, even experts with years of training. Daniel Kahneman explains that this is because human predictions tend to be "noisy": given the same input, different people (or even the same person at different times) will make vastly different predictions. The use of statistical formulas takes the noise out.

<https://medium.com/@natematias/bias-and-noise-daniel-kahneman-on-errors-in-decision-making-6bc844ff5194>

Harms of AI for predicting social outcomes

- Hunger for personal data
- Massive transfer of power from domain experts & workers to unaccountable tech companies
- Lack of explainability
- Distracts from interventions
- Veneer of accuracy
- ...

Compared to manual scoring rules, the use of AI for prediction has many drawbacks.

Perhaps the most significant is the lack of explainability. Instead of points on a driver's license, imagine a system in which every time you get pulled over, the police officer enters your data into a computer. Most times you get to go free, but at some point the black box system tells you you're no longer allowed to drive.

Unfortunately we actually have this kind of system today in many domains.

Takeaways

AI excels at some tasks, but can't predict social outcomes.

We must resist the enormous commercial interests that aim to obfuscate this fact.

In most cases, manual scoring rules are just as accurate, far more transparent, and worth considering.