Can AI be trusted?

Prof David Wong

Declarations

Research Project was funded by Engineering and Physical Science Research Council awarded to the University of Liverpool

AI-sight Ltd is a spin-out company

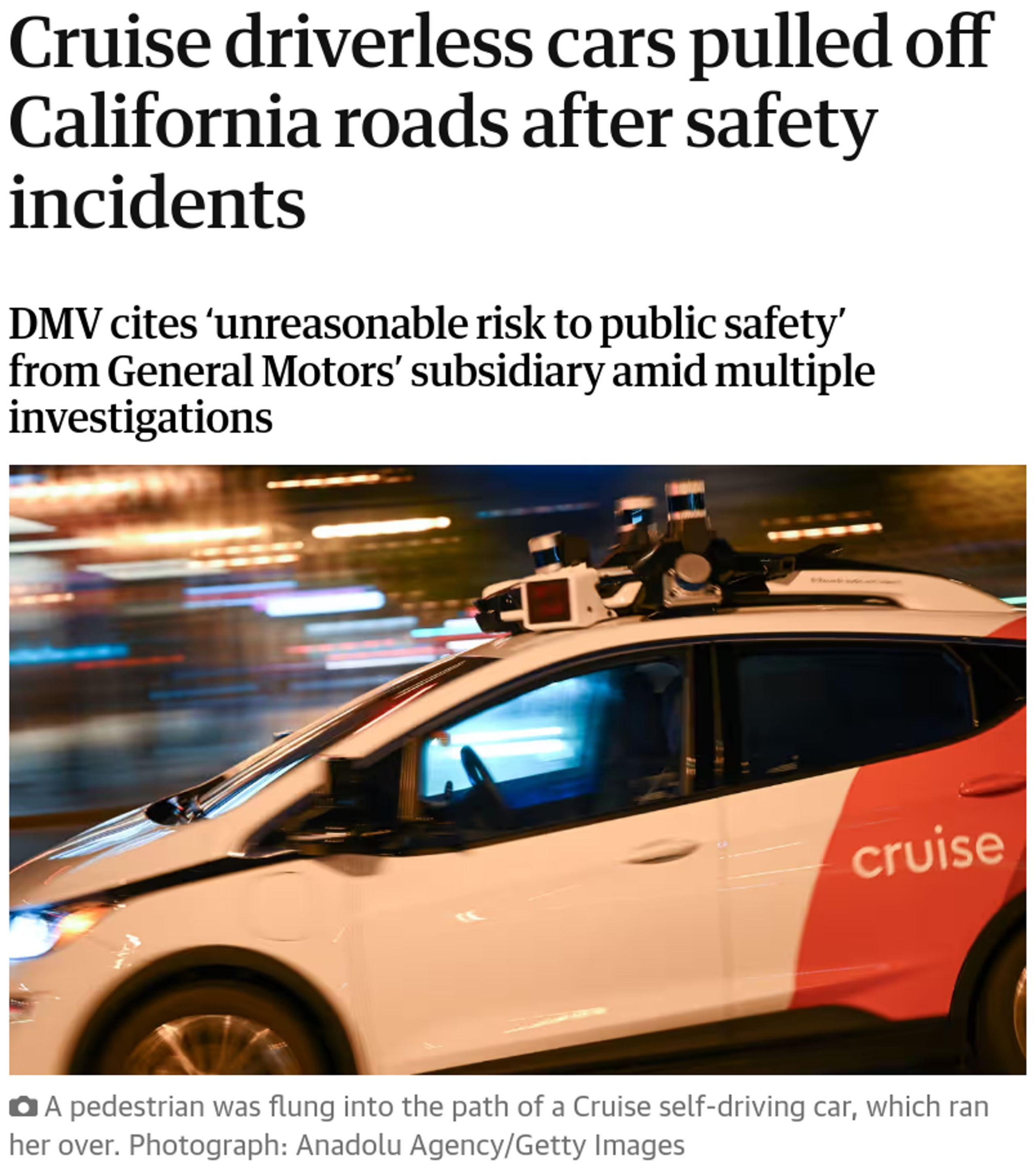
AI-sight
Retinal Health Screening

EPSRC
Engineering and Physical Sciences
Research Council

FIAT LUX
UNIVERSITY OF
LIVERPOOL

Summary

1. Medical test paradox
2. Bayes theorem
3. Gambler’s fallacy
4. Weber’s Law
5. Ranking
6. Fast and slow thinking



Cruise driverless cars pulled off California roads after safety incidents

DMV cites ‘unreasonable risk to public safety’ from General Motors’ subsidiary amid multiple investigations

A pedestrian was flung into the path of a Cruise self-driving car, which ran her over. Photograph: Anadolu Agency/Getty Images

1. Medical test paradox example

A 40 year old man had a routine screening test for testicular cancer. He was told that the test was 99% sensitive and 99% specific for this cancer which has a prevalence of 1 in 10,000. Worried he went and ask his GP, “what are my chances of having cancer as the test is positive?”

- 99%
- 90%?
- 9%?
- 1%?

Most doctors would say high because sensitivity and specificity,

They'd be wrong. Thus it is sometimes called a medical paradox.

Prevalence can simply be expressed a ratio

$$\frac{\text{All patients with disease , }\mathbf{D+}}{\text{All patients without disease, }\mathbf{D-}} \longrightarrow \frac{1}{9{,}999}$$

Bayes Factor or likelihood ratio for a positive test

$$\frac{\text{True positive rate}}{\text{False positive rate}} = \frac{\text{Sensitivity}}{\text{1-specificity}} = \frac{99\%}{1\%}$$

2. Bayes theorem says, given the test is positive, the chance of having the disease

$$\frac{\text{All patients with disease}}{\text{All patients without disease}} \times \frac{\text{TP rate}}{\text{FP rate}}$$

Bayes theorem says, given the test is positive, the chance of having the disease

$$\frac{1}{9{,}999} \times \boxed{\frac{99\%}{1\%}} = \frac{1}{101}$$

Bayes factor or Positive Likelihood Ratio is 99

Is it a good test then? Misleading?

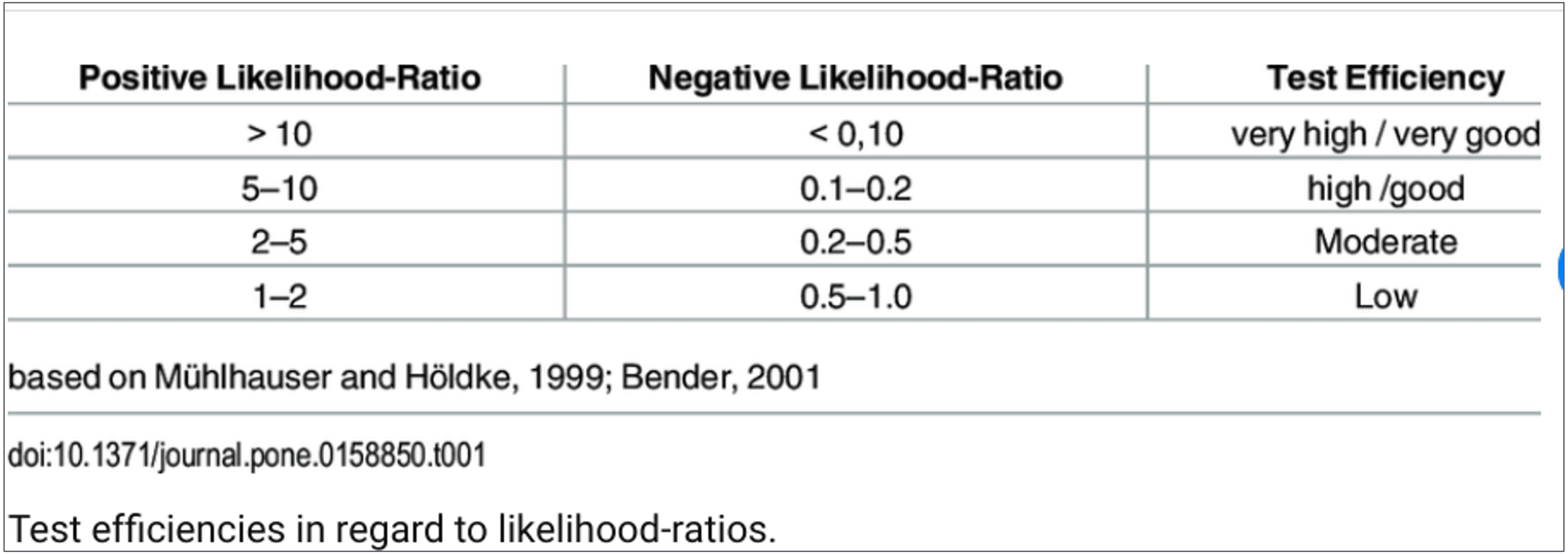
$$\frac{1}{9{,}999} \times \frac{99\%}{1\%} = \frac{1}{101}$$

WRONG!

Concentrated the odds 99 x

Concentrates

Dilutes



Positive Likelihood-Ratio	Negative Likelihood-Ratio	Test Efficiency
> 10	< 0,10	very high / very good
5–10	0.1–0.2	high /good
2–5	0.2–0.5	Moderate
1–2	0.5–1.0	Low

based on Mühlhauser and Höldke, 1999; Bender, 2001

doi:10.1371/journal.pone.0158850.t001

Test efficiencies in regard to likelihood-ratios.

Rule-in and rule-out

SPIN: high specific means high chance of having disease if test positive, i.e. to rule-in disease

SNOUT: high sensitivity means high chance of not having disease if test was negative, i.e. to rule-out disease

Wrong because we always need to know the prevalence

Instead we should use

Bayes Factor (also called likelihood ratios)

Positive Bayes Factor is True positive rate / False positive rate - use this for ruling in

Negative Bayes Factor is False Negative rate / True Negative rate - use this for rule out

In this case, what is the chance of not having the disease if the test was negative

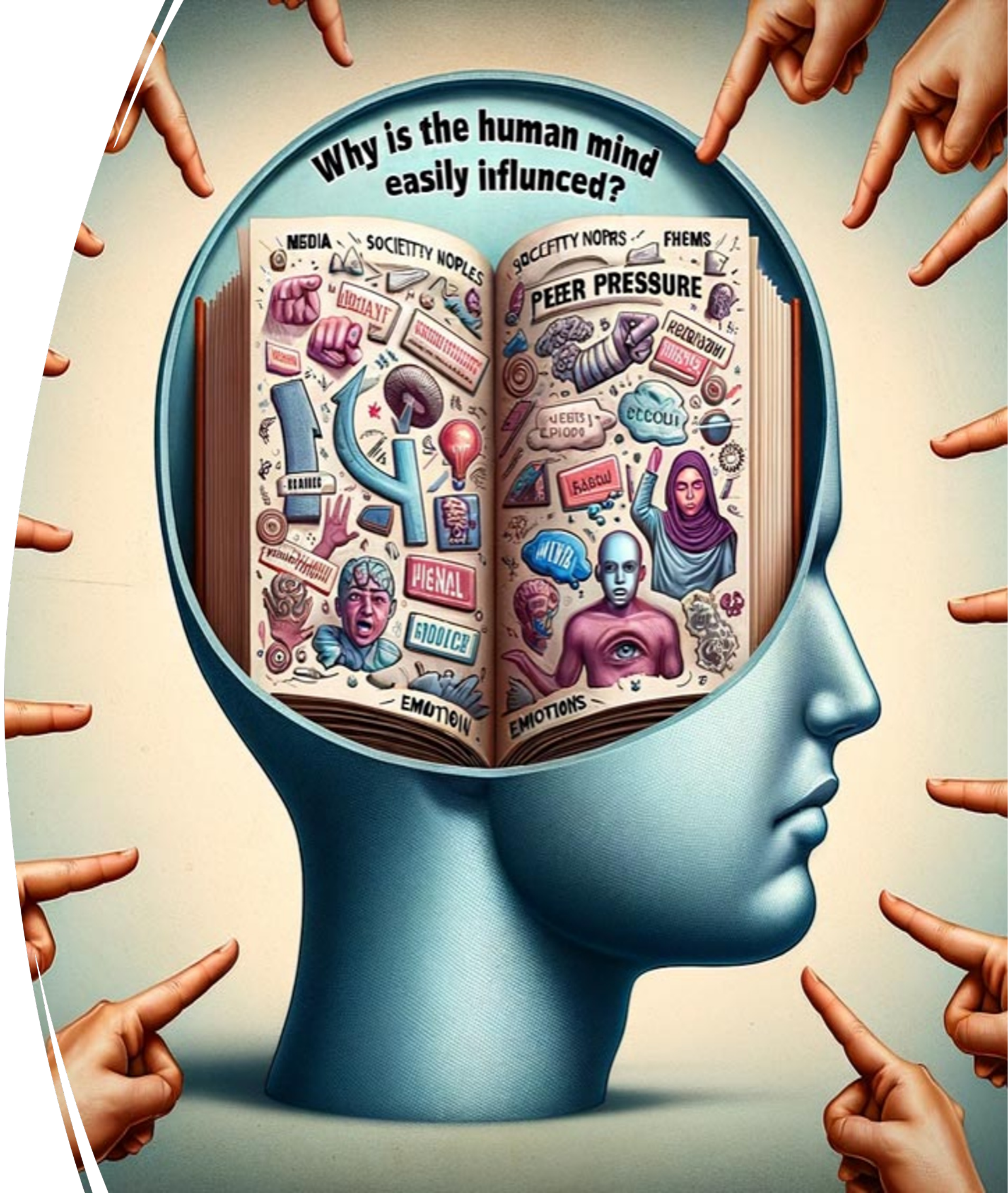
$$\frac{\text{D+ (1)}}{\text{D- (9999)}} \times \frac{\text{False Negative rate (1-sen) (1\%)}}{\text{True Negative rate (spe)} \quad 99\%}$$

Negative 1 chance in 989,901 (diluted 99 x)

AI is trained by humans

Humans are inconsistent and susceptible to biases

Training data is not perfect



In a reading or grading centres

There are typically 5 - 10% significant differences requiring arbitration

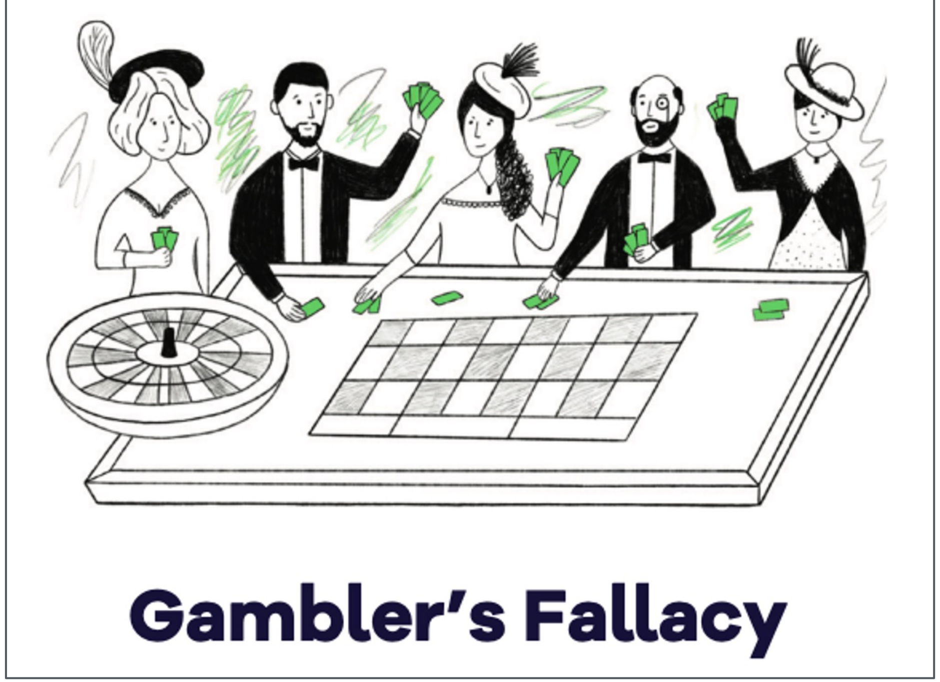
Graders’ thresholds drift and periodic audit exercises are used to align standards.

Why drift?

Images are random

Prevalence is unpredictable

3. Gambler's Fallacy



On Aug 18th, 1913, hordes of gamblers lost millions at the Monte Carlo Casino because "Black" came up 26 times in a row.

How sure are you that your threshold has not drifted?

Working on your own for weeks at a time…

If you graded R2 in 26 consecutive times, would you question your own judgement. Am I referring too many false positives?

We like to believe we are as constant as the Northern Star

Until you compare yourself with an AI?

Could the other person be an AI?

Everything that we have talk about applies to you

Because, you are the test..

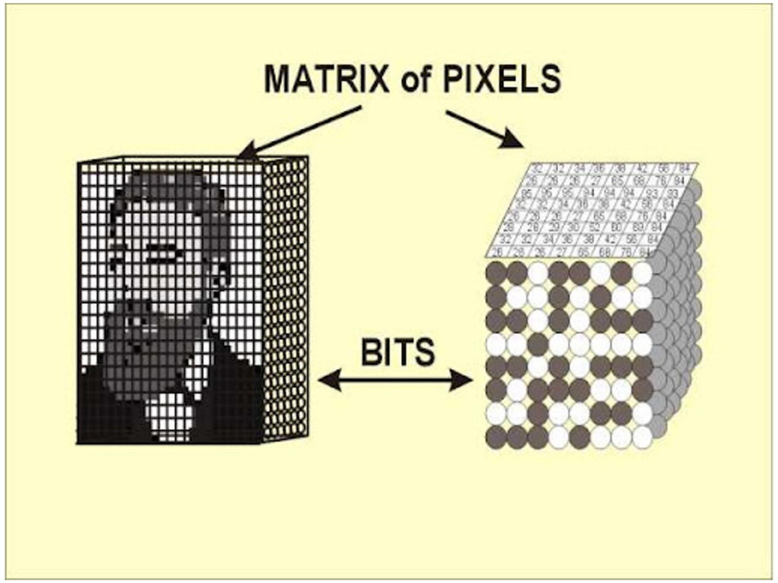
You are also the gold standard

We ask how trustworthy is an AI

But AI and humans “see” differently?

Where humans see features

AI sees numbers

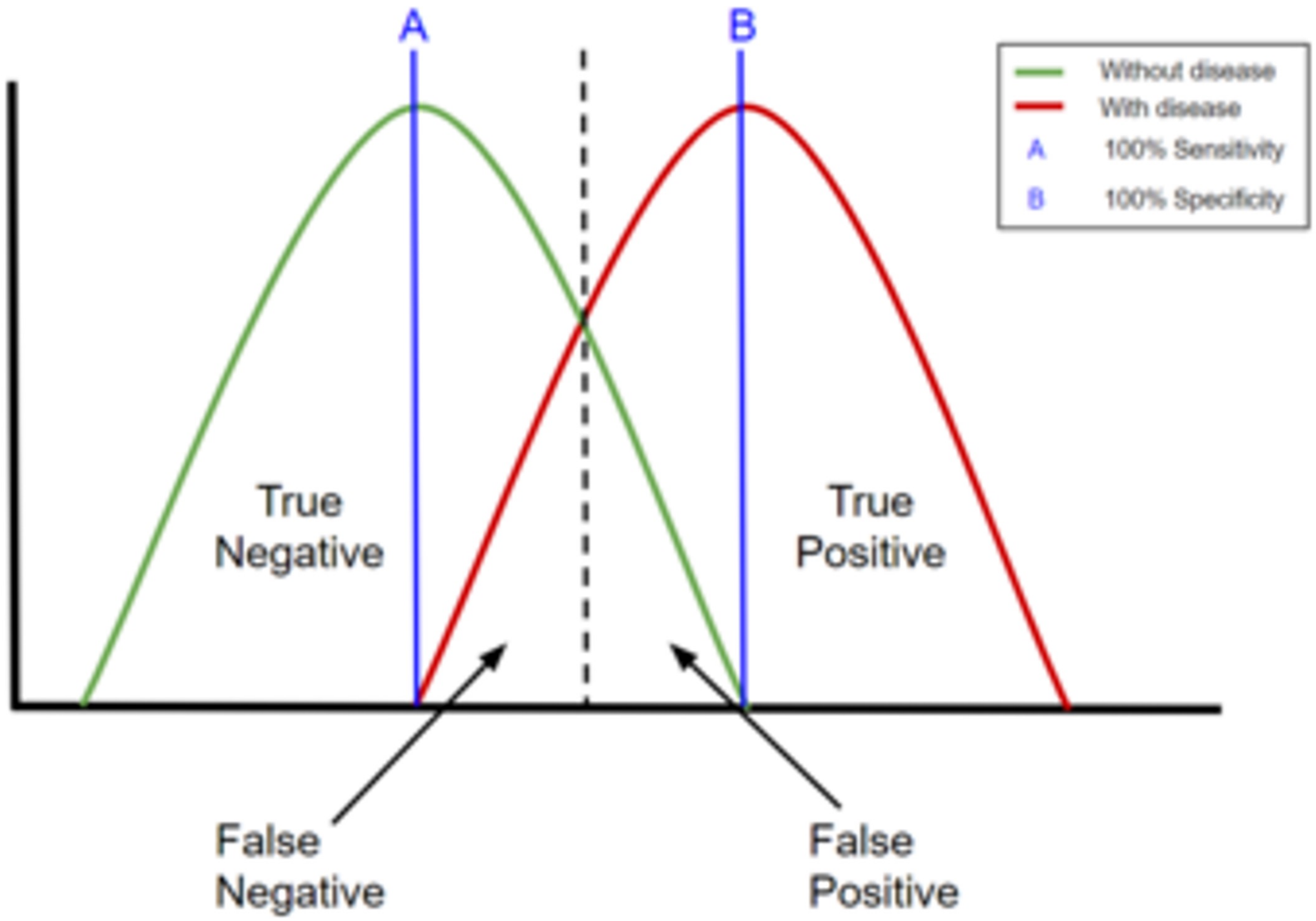


With recognition, there is always of risk

Of mistaken identity or not recognising something you have not seen before

Is it safe?

Sensitivity vs. Specificity



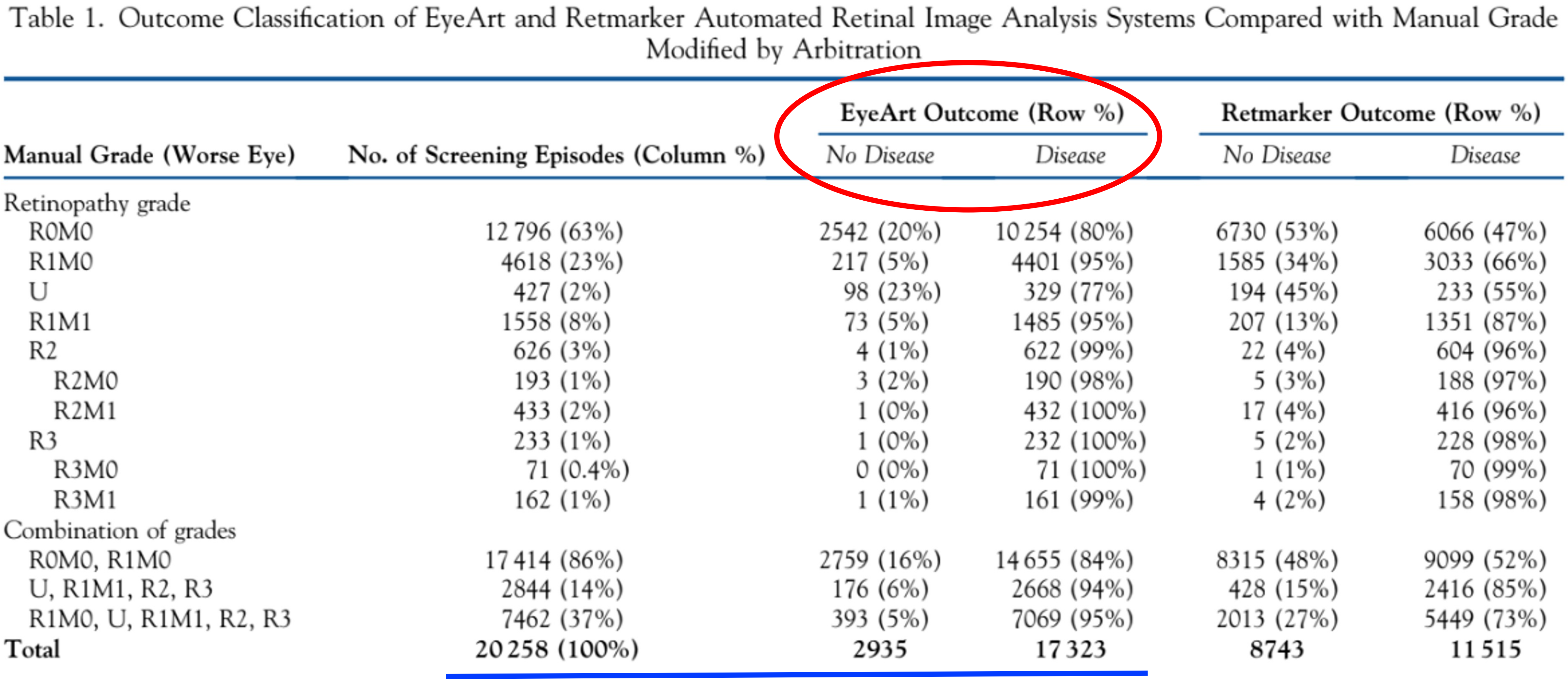


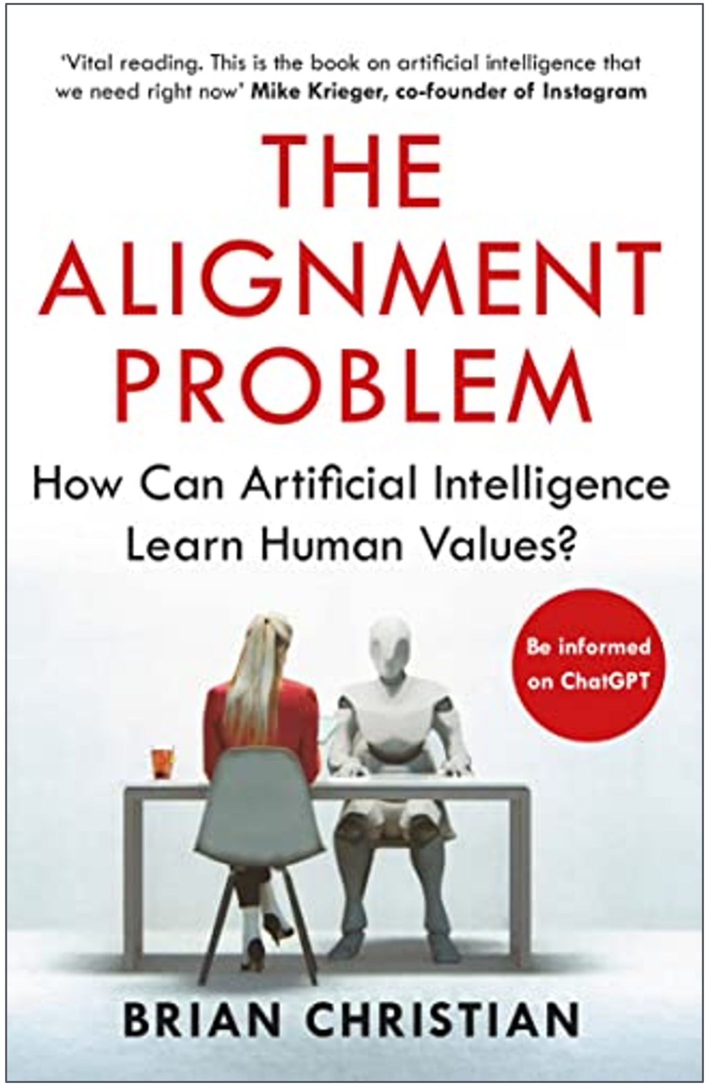
Table 1. Outcome Classification of EyeArt and Retmarker Automated Retinal Image Analysis Systems Compared with Manual Grade Modified by Arbitration

		EyeArt Outcome (Row %)		Retmarker Outcome (Row %)	
Manual Grade (Worse Eye)	**No. of Screening Episodes (Column %)**	*No Disease*	*Disease*	*No Disease*	*Disease*
Retinopathy grade					
R0M0	12 796 (63%)	2542 (20%)	10 254 (80%)	6730 (53%)	6066 (47%)
R1M0	4618 (23%)	217 (5%)	4401 (95%)	1585 (34%)	3033 (66%)
U	427 (2%)	98 (23%)	329 (77%)	194 (45%)	233 (55%)
R1M1	1558 (8%)	73 (5%)	1485 (95%)	207 (13%)	1351 (87%)
R2	626 (3%)	4 (1%)	622 (99%)	22 (4%)	604 (96%)
R2M0	193 (1%)	3 (2%)	190 (98%)	5 (3%)	188 (97%)
R2M1	433 (2%)	1 (0%)	432 (100%)	17 (4%)	416 (96%)
R3	233 (1%)	1 (0%)	232 (100%)	5 (2%)	228 (98%)
R3M0	71 (0.4%)	0 (0%)	71 (100%)	1 (1%)	70 (99%)
R3M1	162 (1%)	1 (1%)	161 (99%)	4 (2%)	158 (98%)
Combination of grades					
R0M0, R1M0	17 414 (86%)	2759 (16%)	14 655 (84%)	8315 (48%)	9099 (52%)
U, R1M1, R2, R3	2844 (14%)	176 (6%)	2668 (94%)	428 (15%)	2416 (85%)
R1M0, U, R1M1, R2, R3	7462 (37%)	393 (5%)	7069 (95%)	2013 (27%)	5449 (73%)
Total	**20 258 (100%)**	**2935**	**17 323**	**8743**	**11 515**

We don’t trust AI

Not just because AI makes errors

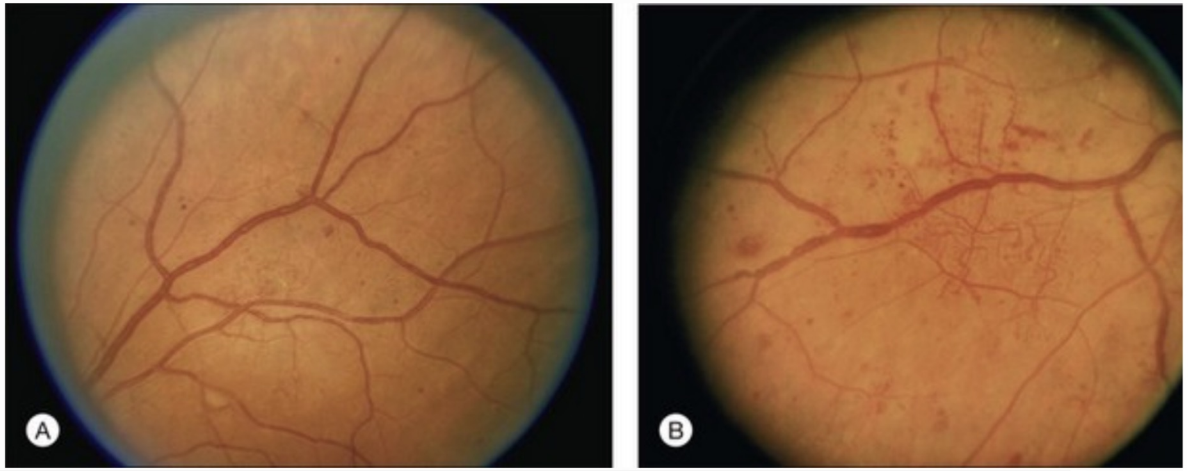
But because they seem unpredictable


'Vital reading. This is the book on artificial intelligence that we need right now' Mike Krieger, co-founder of Instagram
THE ALIGNMENT PROBLEM
How Can Artificial Intelligence Learn Human Values?
Be informed on ChatGPT
BRIAN CHRISTIAN

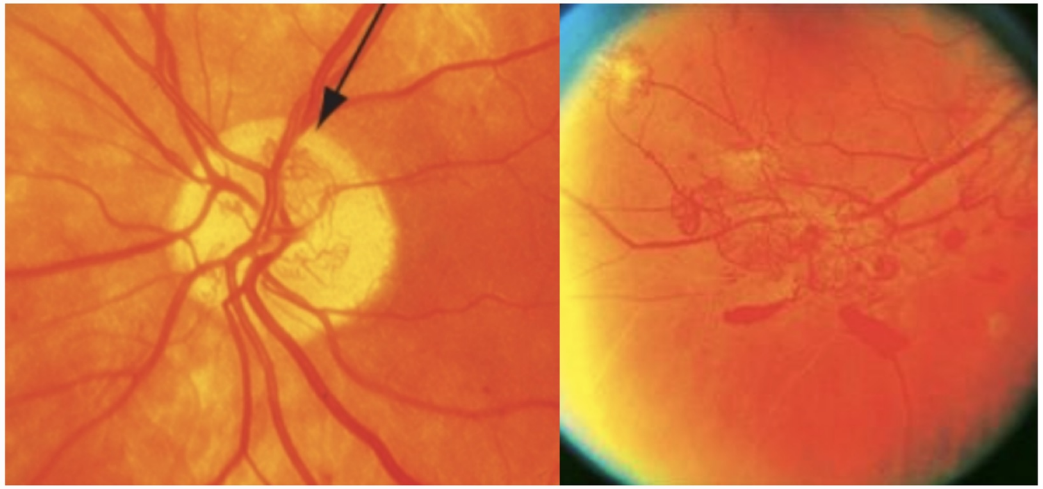
The uncertainty of AI is basically an alignment problem

Can AI learn to think like humans?

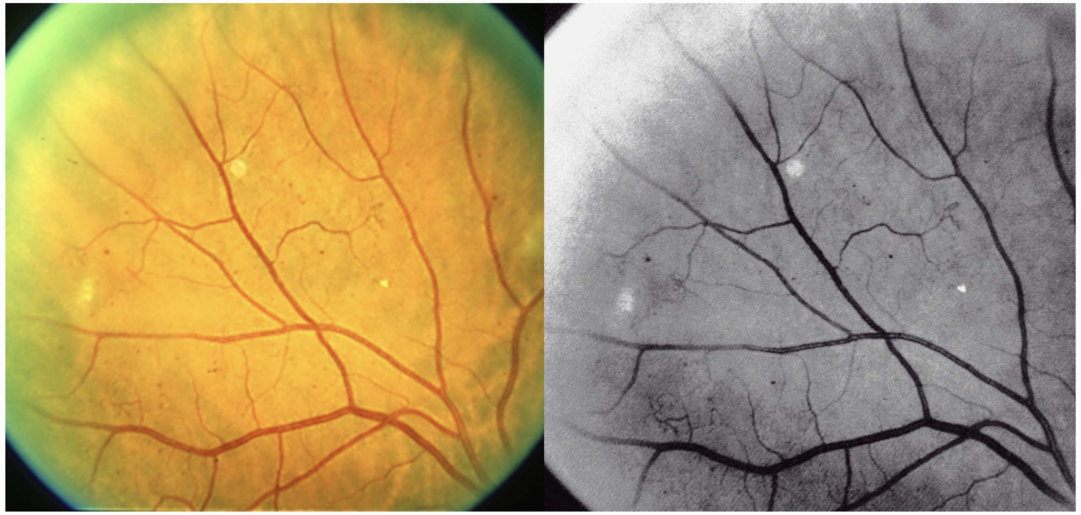
Humans grade by comparison against standards and heuristics



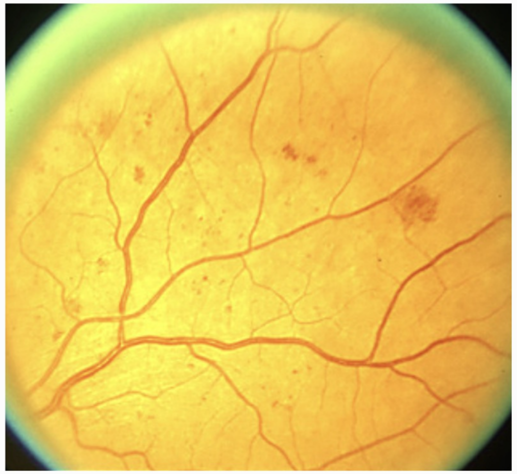
Airlie Hou



Standard photos 10A (left) and 7 (right) depicting approximately 1/3 disc area of NVD and at least 1/2 disc area of NVE, both of which w classify as PDR whether or not vitreous or pre-retinal heme was present.



Standard photo 8A showing the presence of intraretinal microvascular abnormalities. Also note the cotton wool spots.



Standard Photo 2A showing intraretinal (dot-blot) hemorrhages and microaneurysms.

4. The psychophysics of human judgement based on comparison is the Weber law

Our ability to differentiate between 2 items depends on the differences in their quantities.

Therefore, **disagreements between humans inherently tend to be at the boundary or threshold**

Funded by EPSRC, Liverpool developed an AI for DR screening based on Adaptive Comparative Judgement

ACJ

Based on pairwise comparisons,

Given a pair of fundal images, AI is taught which one of the two has the more severe retinopathy

What does ACJ do?

ACJ convert the results of pairwise comparisons into ranking

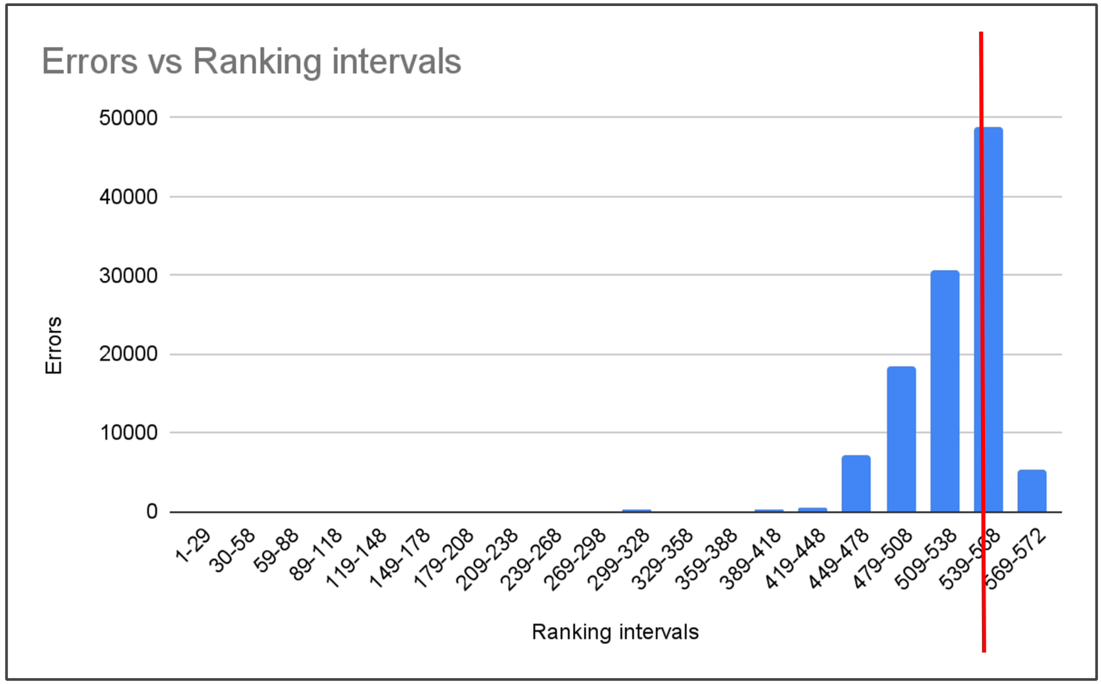
For example, given images, it will rank them in order of severity from 1 to 100.

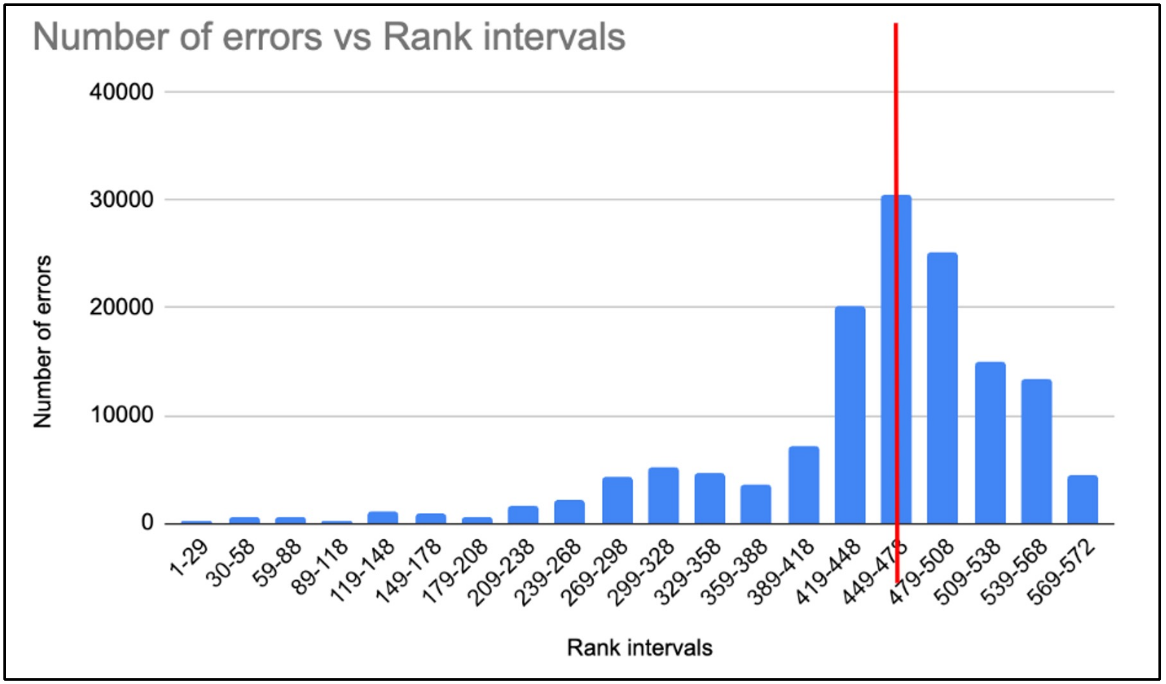
5. Ranking. How does it classify?

If for example you rank the exam results of 40 students from highest to lowest, you can arbitrarily say the bottom 10 failed and top 30 students passed.

In case of the retina, disease / no disease

Does AI based on ACJ obey Weber's law?





Moderate v Severe NPDR

Mild v Moderate NPDR

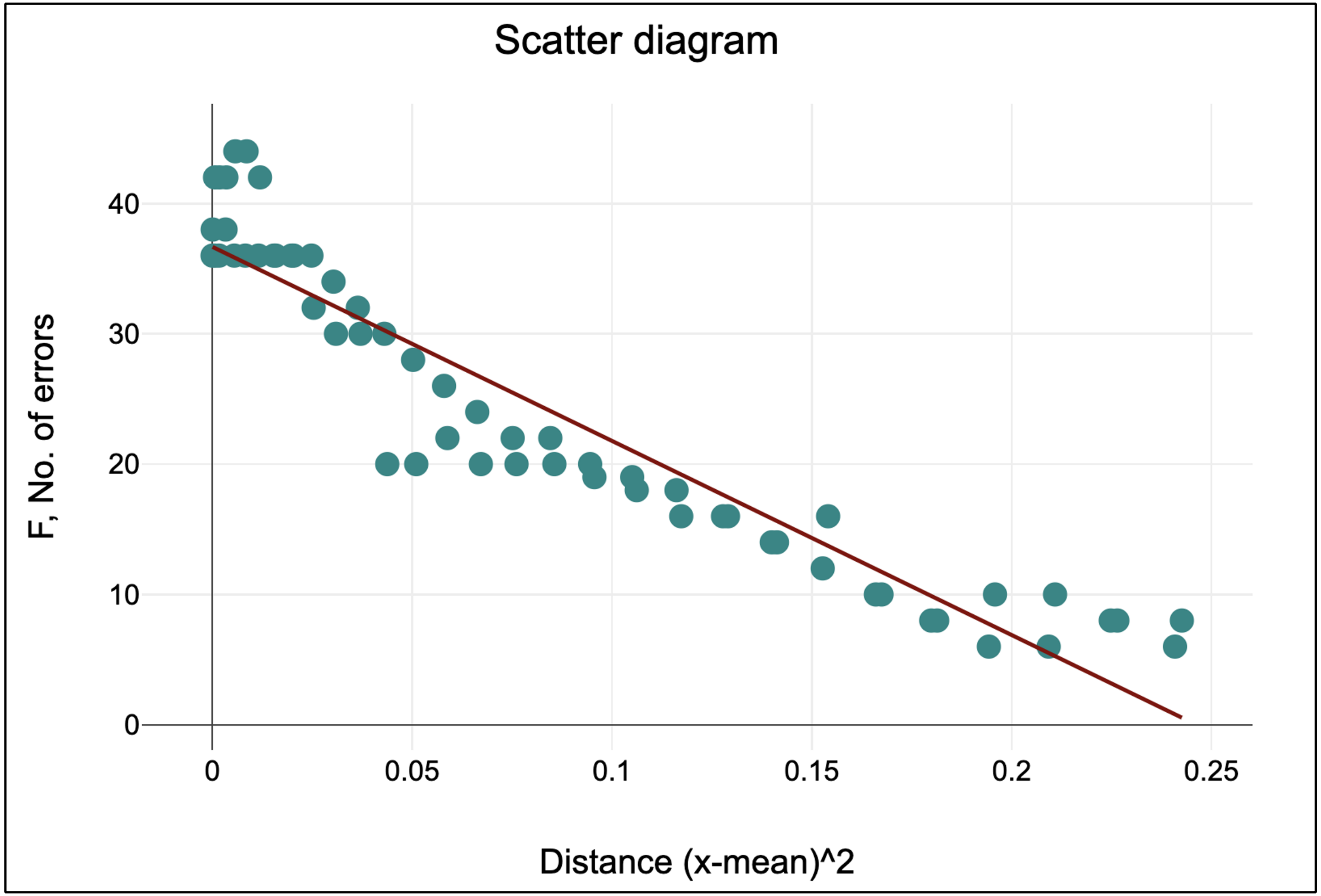
Could it be just chance?

Human ranked 158 images in order of severity

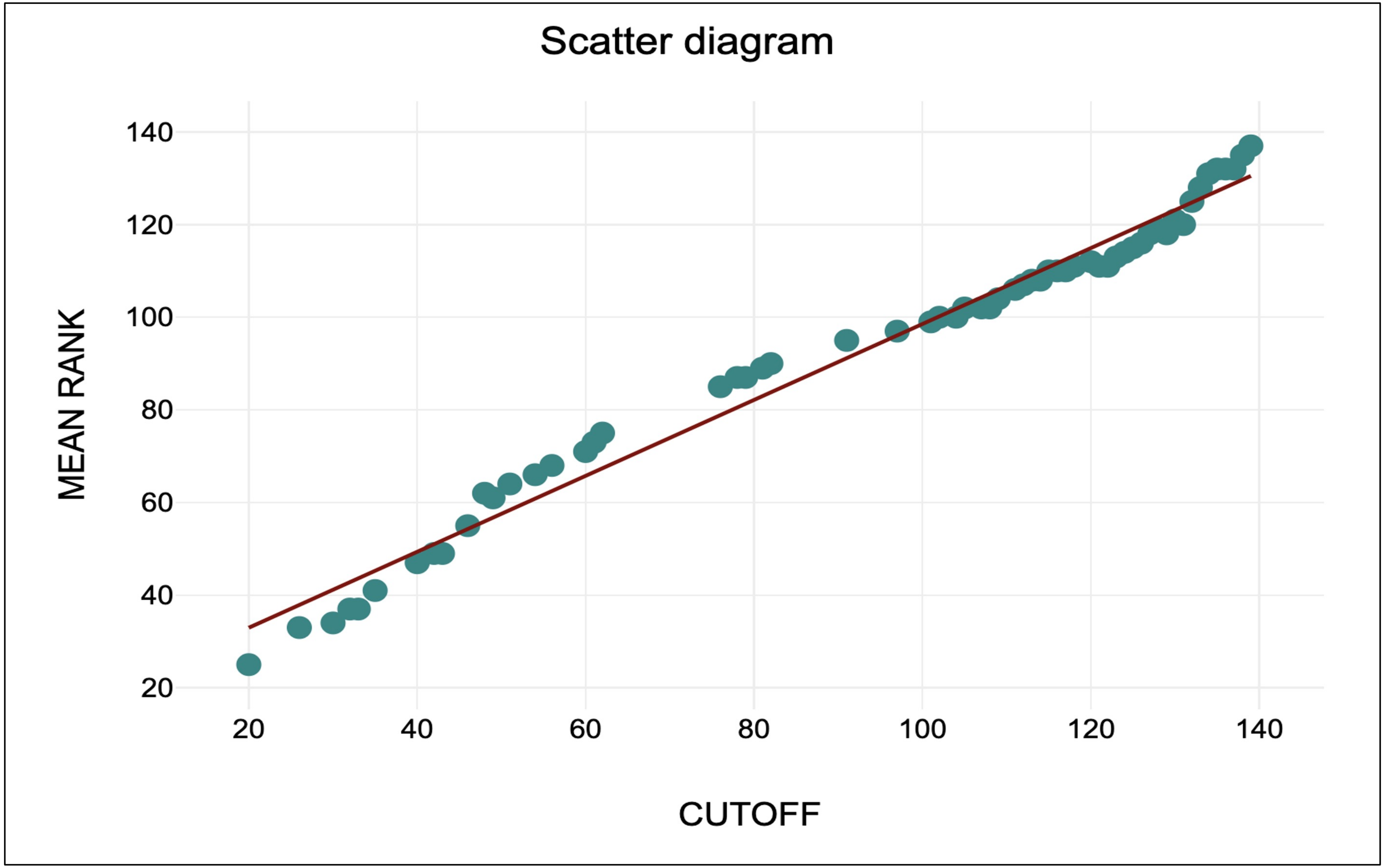
Ranking by 3 experts (TC, SH, DS) masked to AI ranking

61 grades of severity

i.e. 60 boundaries or cutoffs



This plot shows the number of errors against the distance between cutoff and the distance between the images. The Pearson correlation is -0.94 and p<0.001, n=60.

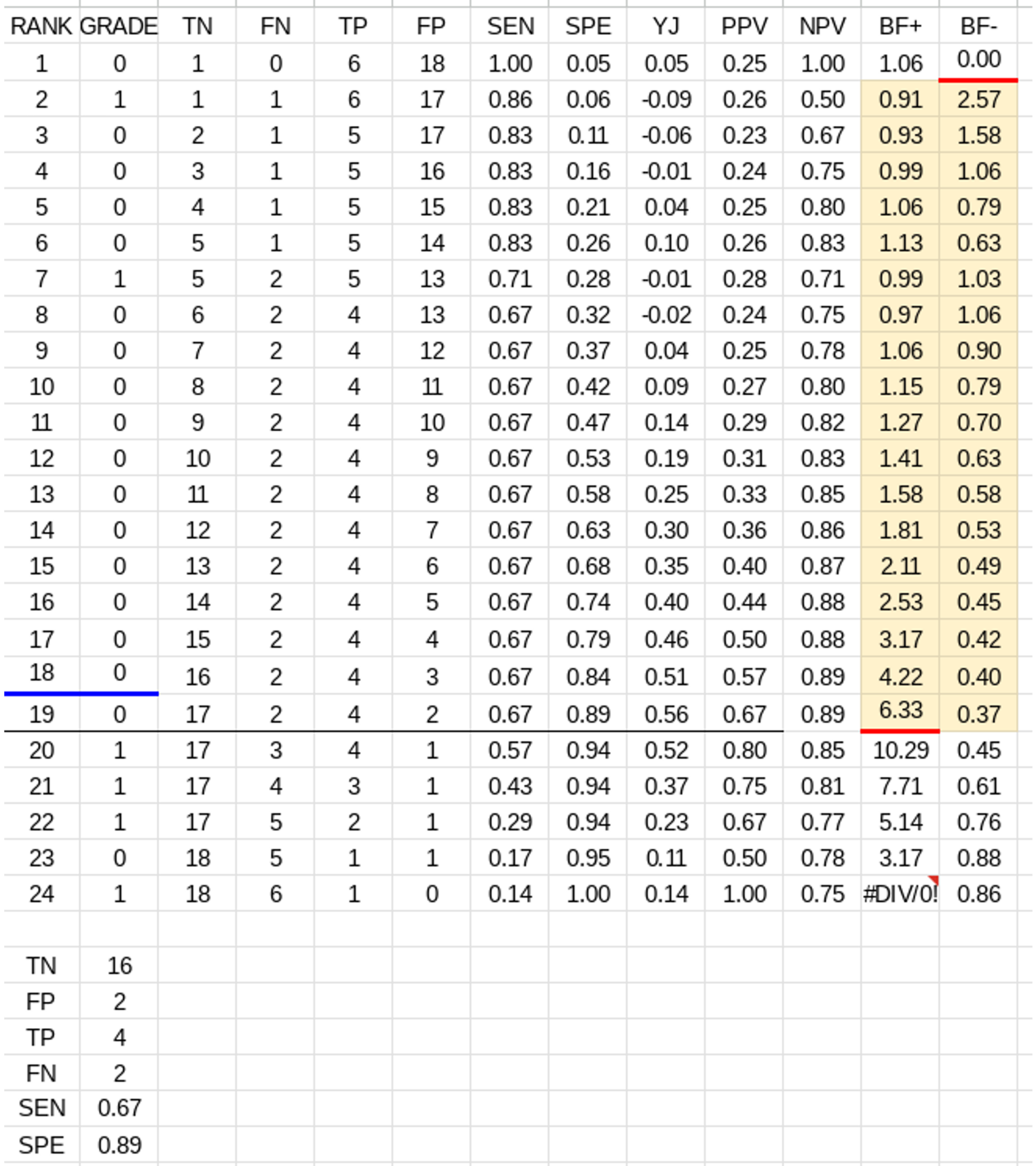


The mean ranks of the distribution of errors were plotted against the 60 cutoffs. We use Pearson to test the correlation. The r was 0.99, the p<0.001, n=61.

Why is clustering so important? Consider this example:

24 cases, 18 normal and 6 abnormal

4 cases misdiagnosed



RANK	GRADE	TN	FN	TP	FP	SEN	SPE	YJ	PPV	NPV	BF+	BF-
1	0	1	0	6	18	1.00	0.05	0.05	0.25	1.00	1.06	0.00
2	1	1	1	6	17	0.86	0.06	-0.09	0.26	0.50	0.91	2.57
3	0	2	1	5	17	0.83	0.11	-0.06	0.23	0.67	0.93	1.58
4	0	3	1	5	16	0.83	0.16	-0.01	0.24	0.75	0.99	1.06
5	0	4	1	5	15	0.83	0.21	0.04	0.25	0.80	1.06	0.79
6	0	5	1	5	14	0.83	0.26	0.10	0.26	0.83	1.13	0.63
7	1	5	2	5	13	0.71	0.28	-0.01	0.28	0.71	0.99	1.03
8	0	6	2	4	13	0.67	0.32	-0.02	0.24	0.75	0.97	1.06
9	0	7	2	4	12	0.67	0.37	0.04	0.25	0.78	1.06	0.90
10	0	8	2	4	11	0.67	0.42	0.09	0.27	0.80	1.15	0.79
11	0	9	2	4	10	0.67	0.47	0.14	0.29	0.82	1.27	0.70
12	0	10	2	4	9	0.67	0.53	0.19	0.31	0.83	1.41	0.63
13	0	11	2	4	8	0.67	0.58	0.25	0.33	0.85	1.58	0.58
14	0	12	2	4	7	0.67	0.63	0.30	0.36	0.86	1.81	0.53
15	0	13	2	4	6	0.67	0.68	0.35	0.40	0.87	2.11	0.49
16	0	14	2	4	5	0.67	0.74	0.40	0.44	0.88	2.53	0.45
17	0	15	2	4	4	0.67	0.79	0.46	0.50	0.88	3.17	0.42
18	0	16	2	4	3	0.67	0.84	0.51	0.57	0.89	4.22	0.40
19	0	17	2	4	2	0.67	0.89	0.56	0.67	0.89	6.33	0.37
20	1	17	3	4	1	0.57	0.94	0.52	0.80	0.85	10.29	0.45
21	1	17	4	3	1	0.43	0.94	0.37	0.75	0.81	7.71	0.61
22	1	17	5	2	1	0.29	0.94	0.23	0.67	0.77	5.14	0.76
23	0	18	5	1	1	0.17	0.95	0.11	0.50	0.78	3.17	0.88
24	1	18	6	1	0	0.14	1.00	0.14	1.00	0.75	#DIV/0!	0.86

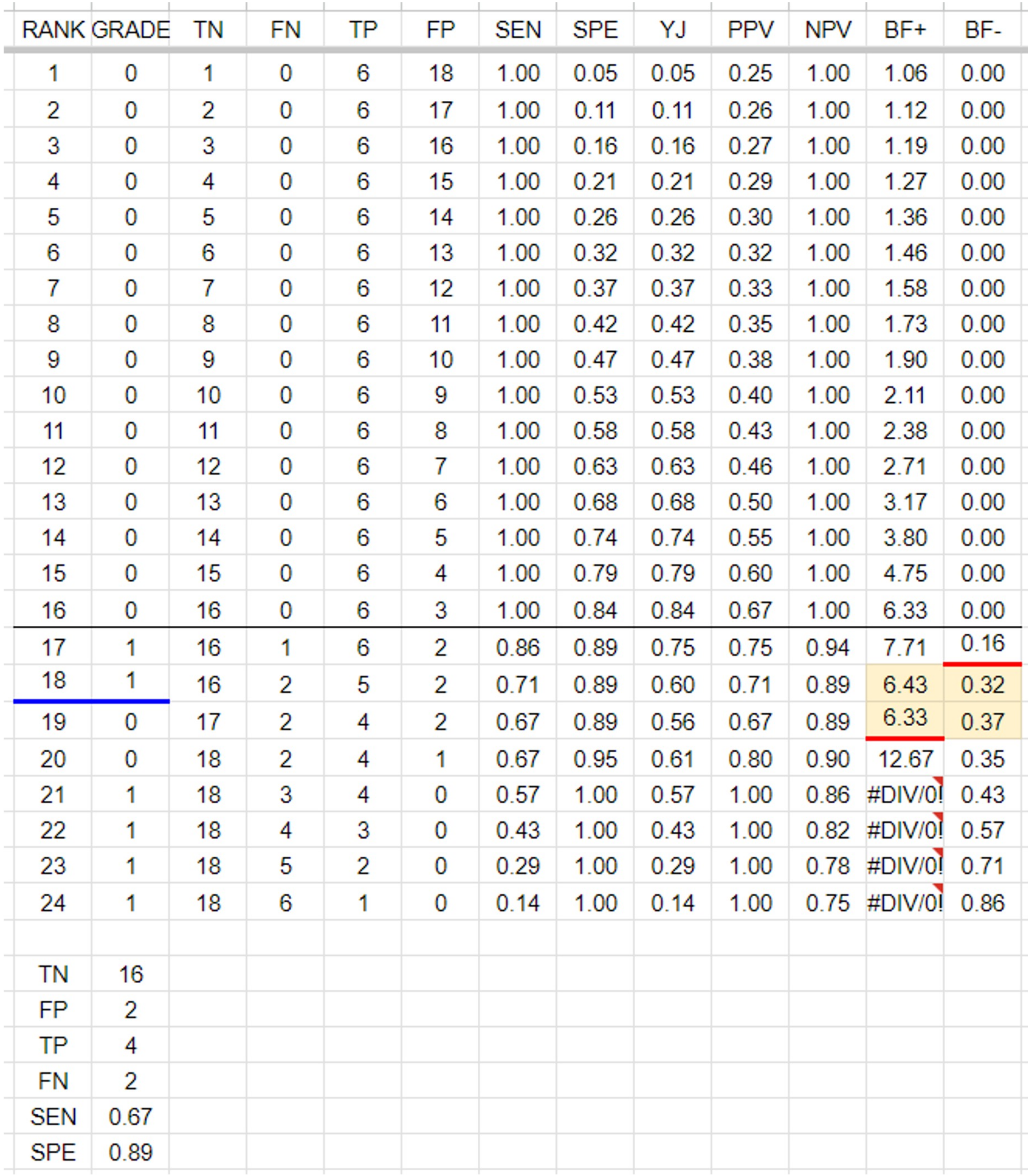
TN	16
FP	2
TP	4
FN	2
SEN	0.67
SPE	0.89

Cut off at first 18

As a test, it is absolutely useless

Blue line marks the prevalence

Sen 0.67 and Spe 0.89



RANK	GRADE	TN	FN	TP	FP	SEN	SPE	YJ	PPV	NPV	BF+	BF-
1	0	1	0	6	18	1.00	0.05	0.05	0.25	1.00	1.06	0.00
2	0	2	0	6	17	1.00	0.11	0.11	0.26	1.00	1.12	0.00
3	0	3	0	6	16	1.00	0.16	0.16	0.27	1.00	1.19	0.00
4	0	4	0	6	15	1.00	0.21	0.21	0.29	1.00	1.27	0.00
5	0	5	0	6	14	1.00	0.26	0.26	0.30	1.00	1.36	0.00
6	0	6	0	6	13	1.00	0.32	0.32	0.32	1.00	1.46	0.00
7	0	7	0	6	12	1.00	0.37	0.37	0.33	1.00	1.58	0.00
8	0	8	0	6	11	1.00	0.42	0.42	0.35	1.00	1.73	0.00
9	0	9	0	6	10	1.00	0.47	0.47	0.38	1.00	1.90	0.00
10	0	10	0	6	9	1.00	0.53	0.53	0.40	1.00	2.11	0.00
11	0	11	0	6	8	1.00	0.58	0.58	0.43	1.00	2.38	0.00
12	0	12	0	6	7	1.00	0.63	0.63	0.46	1.00	2.71	0.00
13	0	13	0	6	6	1.00	0.68	0.68	0.50	1.00	3.17	0.00
14	0	14	0	6	5	1.00	0.74	0.74	0.55	1.00	3.80	0.00
15	0	15	0	6	4	1.00	0.79	0.79	0.60	1.00	4.75	0.00
16	0	16	0	6	3	1.00	0.84	0.84	0.67	1.00	6.33	0.00
17	1	16	1	6	2	0.86	0.89	0.75	0.75	0.94	7.71	0.16
18	1	16	2	5	2	0.71	0.89	0.60	0.71	0.89	6.43	0.32
19	0	17	2	4	2	0.67	0.89	0.56	0.67	0.89	6.33	0.37
20	0	18	2	4	1	0.67	0.95	0.61	0.80	0.90	12.67	0.35
21	1	18	3	4	0	0.57	1.00	0.57	1.00	0.86	#DIV/0!	0.43
22	1	18	4	3	0	0.43	1.00	0.43	1.00	0.82	#DIV/0!	0.57
23	1	18	5	2	0	0.29	1.00	0.29	1.00	0.78	#DIV/0!	0.71
24	1	18	6	1	0	0.14	1.00	0.14	1.00	0.75	#DIV/0!	0.86

TN	16
FP	2
TP	4
FN	2
SEN	0.67
SPE	0.89

If the errors are clustered

As a test, it is absolutely wonderful

Adjust the cutoff to first 16

Sen=1 and Spe=0.84

Despite both sets of results getting 4 out of 24 wrong (16.6%)

It matters greatly where those 2 FN mistakes are situated for ruling out disease.

When sacrificing specificity to gain sensitivity, the number of FN is greatly reduced

Ranking can also eliminate the Gambler’s fallacy

Instead of waiting for the number 11 bus

AI ranks the images in severity

Working alongside an AI can give instant feedback to consolidate your own standards

Predictability is a game changer

Because this could mean a self-aware AI, knowing when it aligns with human

And its own limitation and divert the uncertain images to human

AI can help humans to be more efficient

Not by making humans grading more and images

Instead grading fewer, focusing on the ones that matter

Difficult ones are not

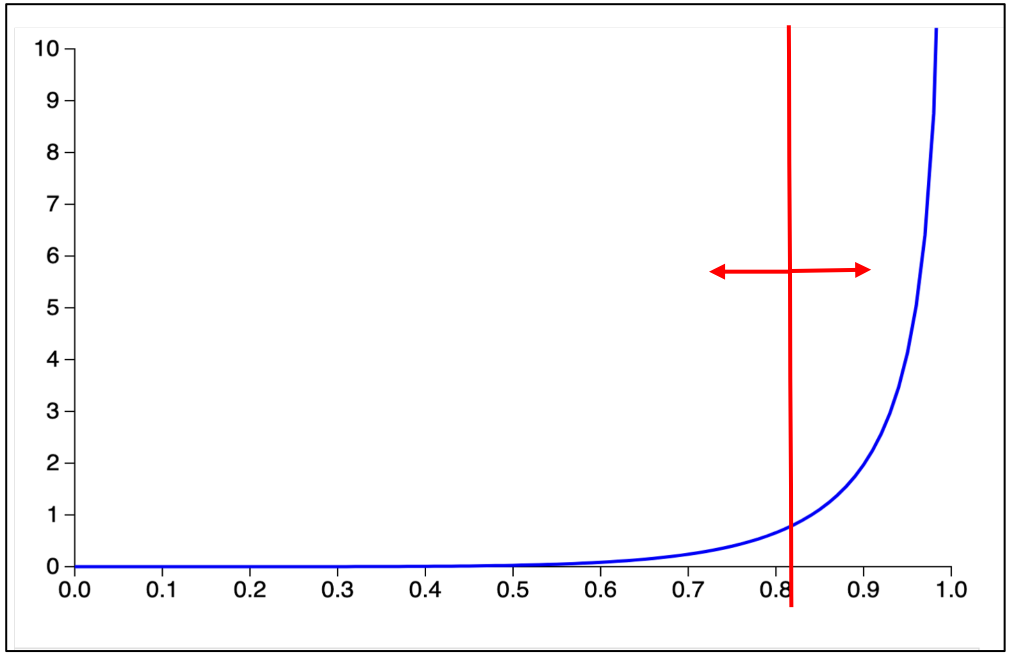
- The normals: double checking 10% of normals don't make sense as by definition you only had a 1 in 10 chance of catching FN if any
- Obviously abnormals: these can be referred directly

It is the borderline cases that needs human input the most

Rank is a proxy for risk

Start from the middle and work outwards

% of error – missing severe NPDR or PDR



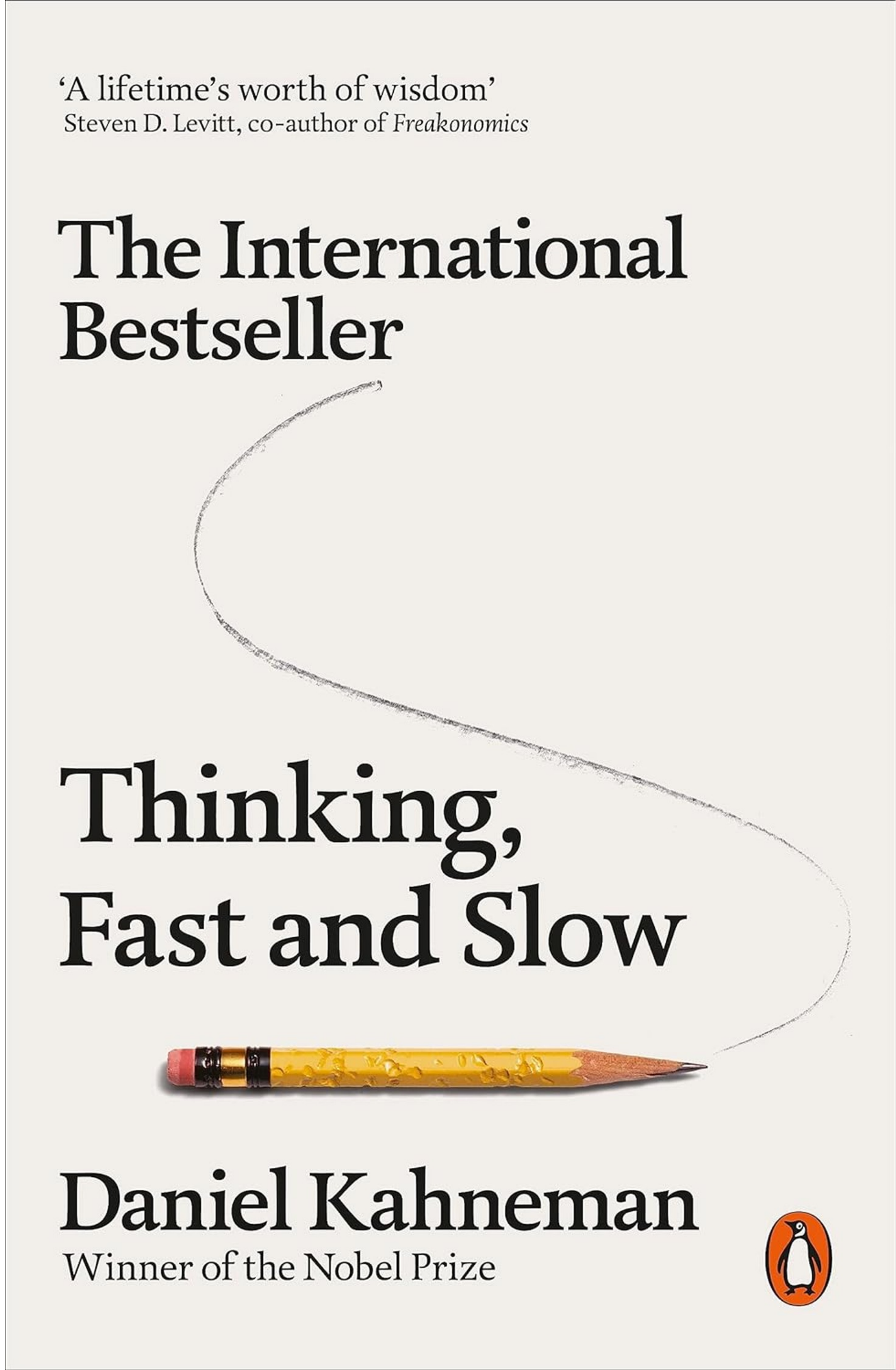
Proportion of cases diagnosed by AI

6. Fast and slow thinking

Do you need to use slow meticulous examination for each images

Fast thinking is often right, less stressful because of cognitive ease

Humans have be better than AI at taking risk



‘A lifetime’s worth of wisdom’
Steven D. Levitt, co-author of *Freakonomics*

The International Bestseller

Thinking, Fast and Slow

Daniel Kahneman
Winner of the Nobel Prize

AI will not and should not replace graders

Lack of expertise (graders) is the biggest barrier to adoption of screening worldwide

If AI is more affordable, more diabetics could be screened, afterall grading is the best example of telemedicine.

There is no reason why UK help the rest of the world

I am not ready for driverless cars

Autopilot for planes are safest