Scoring the Big 3's predictive performance

by Arb Research

- We wanted to know how the old futurists did at forecasting technologies. The client selected the Big Three scifi writers as exemplars with large nonfiction corpuses, not primarily known for their forecasts, thus not cherry-picked.

- We sought a representative sample. We searched systematically using ISFDB, simple pattern-matching, and crowdsourced labour. We ended up checking ~one third of their entire nonfiction.

- We tagged and labelled them with a subjective system, but decomposed into clean quantities (correctness, ex-post difficulty, closeness to pure tech). We're offering a bug bounty for errors. *Outputs*: Asimov file, Heinlein file, Clarke file.

- We introduced a simple score: relevance to tech forecasting × ex-post smartness. Asimov is on top by some margin, but all of them average <20% of the max score. Each of them beat Kurzweil on long-range accuracy.

- We also looked at their "impressiveness / embarrassment". 3 headline results:

	Ratio of Impressive to Embarrassing predictions	Strict tech accuracy [1]	Average score [2]
Asimov	1.9 : 1	57%	22%
Heinlein	1 : 1	36%	11%
Clarke	0.8 : 1	48%	14%

[1] Dated prediction, unambiguously correct. Taking ambiguous rows out of the denominator. Doesn't account for differences in difficulty!
[2] Where 100% is being invariably correct about difficult technologies, 100 years out. Does account for difficulty.

> *I do not for a moment suggest that more than 1 per cent of science-fiction readers would be reliable prophets; but I do suggest that almost a hundred per cent of reliable prophets will be science-fiction readers – or writers.*

– Clarke (1962)

Intro

When they check old predictions at all, people tend to pick out outstanding wins or amusing misses. We were instead tasked with *systematically* reviewing predictions made by "the Big Three" scifi writers: Arthur C Clarke, Robert Heinlein, and Isaac Asimov. Systematic, as opposed to exhaustive. (See Karnofsky's post for the point of all this.)

Target: Relatively precise predictions about the future course of technology, with a resolution date, in their nonfiction work and interviews. This spec rules out >90% of their writing.

The client asked for relevance-to-EA ("*How well does this forecast fit into the reference class of forecasting far-off future technologies?*") and smartness ("*How smart/dumb does this forecast look in hindsight*?"). Operationalisations are in "Methodology", below.

We also wanted to move fast, delivering within 6 weeks. (Later, 7.5 weeks.)

Initial exploration

I (Gavin) went through the first dozen pages of Google results for queries like "Heinlein predictions", getting mostly fun newspaper or magazine spots. This yielded 200 cherry-picked predictions, and taught me the shape of the problem (for instance, what words are most common in a prediction sentence). I scored them manually, including a crude categorisation of their confidence and so calibration.

Data Collection

We first obtained digital copies of as much of their nonfiction as possible (books, essays, interviews). The resulting intake is 475 files covering ~33% of their nonfiction corpuses, as catalogued by ISFDB and supplemented by a few entries which that database overlooks as insufficiently science-fictional.[3]

Methodology

Judicious use of ocrmypdf and ebook-convert gave us the text without custom ML. Our own pipeline software is also deterministic.

We considered building something semi-supervised for munging the predictions. We ended up using a very simple string-matching regex instead, and then manually inspecting these prediction candidates (about 60,000 sentences). We preserve the verbatim passage and the surrounding context, for easy validation.

This wasn't necessarily the most efficient method (it generated 90% nonpredictions and pseudopredictions), but the variance on how long it would take was low.

We then categorised them and deduplicated by topic; all dropped rows are in the "Duplicates by theme" tab. This was a fairly narrow definition of duplicate - on the order of two rows talking about particular parameters of ion thrust.

The coding is as follows

Correctness:

0 - unambiguously wrong;
1 - ambiguous or near miss;
2 - unambiguously right

Category:

Tech - "does the tech exist at all?". Weighting: 1.
*Tech * econ* - "is the tech cheap / widespread / relevant?". Weighting: 0.66.
*Tech * culture* - "is the tech accepted / widespread / legal?". Weighting: 0.33.
N/A - some other more cultural phenomena. Weighting: 0.

[3] Giving essays and interviews 1/20 the weight of books for the sake of quick comparison.

Relevance to AI:

Category weight * log(years between prediction and resolution)

Difficulty:

1 - was already generally known
2 - was expert consensus
3 - speculative but on trend
4 - above trend, or oddly detailed
5 - prescient, no trend to go off

Smartness:

Correctness * Difficulty

Final Score:

Relevance * Smartness
In practice, as of 2022, the score is bounded in (0,2).

Lastly, we choose what to do with the predictions (a majority) with no stated resolution date. One simple method is to impute the author's average prediction interval and evaluate each prediction assuming that. Asimov tends to predict 68 years out, with std = 32. (This is applied sensibly; we don't impute predictions which say "one day far off", or "eventually". But this is one last added degree of freedom.) In this report, "strict" results are those which ignore these imputed rows.

One difference from the main Cold Takes analysis is that we use all resolved rows (e.g.: if Asimov predicts that we'll have maglev trains by 2030, and we build one in 1994, our analysis uses this row and Holden's doesn't). One problem with this: it's easier to resolve a positive early than a negative, and people tend to predict positives.

Impressiveness / Embarrassment

The score misses something about the spread and wildness of each author. Score is a decent measure of impressiveness though.

We add a measure of embarrassment: how easy the question is, how close in time the prediction was to the resolution.

$$\textit{Embarrassment} = \left(1 / \text{difficulty} + 1 / \log(\text{gap in years}) \right) / 2 \quad \text{if incorrect}$$

We joined these into one “magnitude of notability”. You can see the resulting most notable predictions in Asimov, Heinlein, Clarke.[4]

It’s also interesting to look at how often they were particularly impressive. We binarised the scores as follows:

- "Very impressive": score > 0.9
- "Impressive": 0.5 > score > 0.9
- "Very embarrassing": embarrassment > 1
- "embarrassing": embarrassment > 0.5

See Table 3 for results.

Resulting sample

- Asimov file,
- Heinlein file,
- Clarke file.

	# Predictions	# Resolved predictions	Average difficulty	Average range (years)
Asimov	496	149	2.9 / 5	68
Heinlein	130	79	2.8 / 5	44
Clarke	890	356	2.4 / 5	37

[4] This is just the tech predictions, so we don't get Heinlein's epic 1949 prediction that Chinese communism alone would survive by 2000.

Table 1. The sample of predictions.
For us, most relevant are timed predictions and tech (i.e. first prototype) predictions.

Results

Our script is here.

Relative performance

If we use Kurzweil as a baseline for "good futurism", then the Big Three come off quite well.[5] Their predictions average 40 years out, and their strict accuracy around 36% compares well to Kurzweil's 20 year predictions having 12 – 24% accuracy. (Note that there are many social predictions in the Kurzweil mix.)

Table 2's "Tech score" represents the most relevant subset: predictions about pure tech which have resolved. The "All category score" includes tech questions plus questions with a softer economic or cultural dependence; each of the Big 3 look worse at these. Recall that the practical upper bound for each score is 2.

	Tech score	All category score	Standard deviation	Strict tech accuracy[6]
Asimov	0.74	0.43	0.5	57%
Heinlein	0.33	0.22	0.4	36%
Clarke	0.37	0.28	0.4	48%

*Table 2. Mean score (relevance * smartness) per author*

The Heinlein sample is small (n=10 for the strict conditions). See the sheet for various permutations (including near misses, including predictions about technology economics, including things which haven't reached their imputed date yet). They don't change much.

[5] Herman Kahn might be a better choice.
[6] Dated prediction, unambiguously correct, about a tech prototype. Taking ambiguous rows out of the denominator. This differs from Karnofsky's subset, which does these filters and then also difficulty > 4.

As noted above, the score isn't that satisfying. So we look at how often each author makes an impressive prediction, vs how often they make an embarrassingly wrong prediction (Table 3):

	Ratio of Impressive to Embarrassing predictions	Ratio of Very Impressive to Very Embarrassing
Asimov	1.9 : 1	30 : 1
Heinlein	1 : 1	- [7]
Clarke	0.8 : 1	13 : 1

Table 3. Wildness / spread of each author

Absolute performance

The scale is bounded above by log(100) = 2 (since the earliest predictions were very roughly 100 years ago), so there's some sense that Asimov was 0.55/2 ≈ 28% as smart as possible, and Heinlein 15%. (This is how the % results on page 1 were obtained.)

This project is an incomplete snapshot of their performance. For instance, Asimov liked to talk about farms on the moon – a question which had a relevant update just two weeks ago. In 2084 the last of their dated predictions will resolve. I expect some movement.

Heinlein brags a lot about his "prophecies", but the big ones (waldoes, water beds) are all in deniable fiction form. He's remarkably good on social change - for instance, he predicts that by 2000, communism will be gone, except in China - but poor (and dishonest, rules-lawyering) on tech change.

Validation

We're convenience sampling from the corpus and then applying several noisy filters to that sample. So there are a few places bias could enter our resulting estimates of performance. Here are the ones we checked:

[7] Again, the small size of the Heinlein sample cheats us. Laplace gives him 84:1.

Collection process

Are ebooks suitably representative of the authors' epistemics? Are out-of-print books much worse, for instance? Seems unlikely to be worse.

To be safe, we bought a book which hasn't been digitised (as far as we know), and paid someone to manually extract the predictions. Clarke's "July 19th, 2019".[8]

The book predictions have an average score of 0.29, compared to Clarke's normal 0.21. So weak reason to think that our digitised sample is not unrepresentatively good, compared to an example of Clarke's undigitised work.

Prediction regex

Is our simple matching regex good enough; how much does it miss?
We used the short newspaper pieces manually labelled as predictions from before to check. The regex caught all 49 of the predictions from the Asimov sample, with a dozen false positives. (0% false negative) The piece only has 150 sentences.

A more extensive piece, Asimov's book of 66 essays "On the Past, Present and Future", led to the regex finding 87% of the 191 manually labelled predictions – a 13% false negative rate.

Crowdsourcing step

What about the crowdsourcing step, which is supposed to be ground truth about whether something is a prediction or not? I included the texts I manually labelled in the first crowdsourced batch. Their accuracy was 98%.

Judgments

Our evaluation consists of thousands of partially subjective judgments of difficulty, correctness, and category. As a very partial mitigation of this, each correctness score had double entry from at least two people, sometimes three.

[8] There is a PDF floating around for "Arthur C. Clarke - July 20, 2019_ A Day in the Life of the 21st Century-Grafton (1987)", but it's 90 pages instead of 500.

Bug bounty

This project consists of thousands and thousands of judgment calls atop vague and underspecified sentences. Most judgments are not very ambiguous, but many are.

If you spot something off in the Predictions tab, we'll pay $5 per cell we update as a result. You should be able to comment on the sheet directly. We'll add *all* criticisms – where we agree and update *or* reject it – to this document for transparency.

Limitations

- Fiction is these people's main work, and their novels contain many tacit predictions – probably most of their predictions. But accuracy is never the first goal in their fiction and so we excluded all of these. (There's an element of entertainment even in their nonfiction, but less severe.)

- The predictions are usually very vague. Almost none take the form "By Year X technology Y will pass on metric Z". This is by contrast with e.g. Kurzweil.

- We originally planned to do 4 passes for each label using MTurk, but I wasn't impressed with the results from the pilot, so we ended up doing two independent passes with a small handpicked team instead, with one supervisor checking the pass.

- We ignored updates and counted both directions. (e.g. Asimov drastically changed his view on technological unemployment between 1960 and 1980.)

- We removed duplicate text, but we didn't deduplicate by prediction *topic*. Repeated predictions reflect the importance and confidence of the author, and so duplication is somewhat relevant to our estimate of their judgment.

Possible extensions

- *More books*. It would be fairly costly to get more of the corpus.
- *Closer inspection of the sampled books.* It would be very expensive to be exhaustive about predictions if done by hand, but an NLP system is doable.
- *Research to remove ambiguity*. When there's reasonable ambiguity about the outcome, we coded it as partially correct rather than investing the time to resolve it further. This would be fairly expensive, maybe 10 mins per ambiguity.
- *Calibration*. Adding a crude calibration estimate (Hi / Med / Lo) would be cheap but not very useful.

Misc

- There's a residual left over by the score, obviously: *general good judgment*, which Asimov wins on. They're all pretty obsessed with overpopulation and similar quasi-technological issues. But Asimov is able to say that fusion will not be here 50 years out, and so on. Clarke is happy to write a whole book flirting with cryptids and ghosts, and gets worse as he ages. Heinlein is a ranter.

- Heinlein wrote a bunch of fake dated predictions in a story called "The Third Millennium Opens" (e.g. FTL in 2000). But he believes half of them, as the footnotes make clear. This is annoying.

- Some of the fictional predictions really are good:
    - *Clarke: "[In] Prelude to Space* (written in 1947), I am amused to see that though I scored a direct hit by giving 1959 as the date of the first Moon-rocket, I put manned satellites in 1970 and the landing on the Moon in 1978. This seemed wildly optimitistic to most people at the time, but now demonstrates my innate conservatism. A still better proof of this is provided by the fact that I made no attempt whatsoever, in 1945, to patent the communication satellite. (See Chapter 16.) I couldn't have done so, as it happens; but at least I would have made the effort, had I dreamed that the first experimental models would be operating while I was still in my forties."

- Clarke (1960):

  "We are still decades – but not centuries – from building such a machine [AGI], yet already we are sure that could be done... The fact that the great computers of today are still highspeed morons, capable of doing nothing beyond the scope of the instructions carefully programmed into them, has given many people a spurious sense of security. No machine, they argue, can possibly be more intelligent than its makers – the men who designed it, and planned its functions. It may be a million times faster in operation, but this is quite irrelevant. Anything and everything that an electronic brain can do must also be within the scope of a human brain, if it had sufficient time and patience. Above all, no machine can show originality or creative power or the other attributes which are fondly labelled 'human'. The argument is wholly fallacious; those who still bring it forth are like the buggy-whip makers who used to poke fun at stranded Model T's. Even if it were true, it could give no comfort, as a careful reading of these remarks by Dr Norbert Wiener will show:

  'This attitude (the assumption that machines cannot possess any degree of originality) in my opinion should be rejected entirely ... It is my thesis that machines can and do transcend some of the limitations of their designers ... It may well be that in principle we cannot make any machine, the elements of whose behaviour we cannot comprehend sooner or later. This does not mean in any way that we shall be able to comprehend them in substantially less time than the operation of the machine, nor even within any given number of years or generations ... This means that though they are theoretically subject to human criticism, such criticism may be ineffective until a time long after it is relevant.'

  In other words, even machines less intelligent than men might escape from our control by sheer speed of operation. And in fact, there is every reason to suppose that machines will become much more intelligent than their builders, as well as incomparably faster."

- Many of the predictions from interviews are totally off the cuff, e.g. Letterman goading Asimov about when cross-country phonecalls won't suck. This is good, since it injects a bit of randomness into questions Asimov is answering. But mark them all as low confidence, since entertainment dominates.

- Another thing to think about is their predictions' causal effect on some questions. This is plausible for things where a small number of Western nerds were on the critical path (e.g. many technologies). Probably too hard to quantify this, but we can flag the obvious ones.

- Asimov predicts Hyperloop in 2084. Sounds right.

- One of the most common clauses in Asimov is “if we are to survive”. He’s usually dead wrong about it.

Acknowledgments

Thanks to Anisiia, Kristi, Christian, Calum, and Paul for massive amounts of careful collection and tagging work. Thanks to John Morrice for the pipeline software.

Regex used to filter the corpus for possible predictions:

(year
will
won\'t
going to
can
can\'t
could
couldn\'t
would
wouldn\'t
may
might
must
shall
should
likely
certain
certainly
definitely
possib
probable
probably
improbable
improbably
chance
maybe
perhaps
sure
surely
percent
remote
doubt
undoubtedly
indubitabl
doubtless
doubtful
assuredly

unquestionably
beyond question
undeniably
incontrovertibl
irrefutably
unequivocally
clearly
plainly
obviously
patently
positively
absolutely
decidedly
cannot rule out
cannot dismiss
cannot discount
believe
expect
predict
anticipate
think
figure
suppose
forecast
foretell
foresee
prognosticate
project
speculate
envision
envisage
i believe
estimate
imagine
picture
conjecture
guess
hazard
augur

presage
hence)