Further Reading

The ideas touched here barely scratch the surface of what we think can be done with algebraic evaluators and AI safety in general. This notebook shares books, articles and resources relevant to the problem of evaluating noisy judges.

Online Resources

The list we have chosen here is eclectic. It reflects sources that have inspired the work behind Data Engines. As such, they are not authoritative about covering the topics upon which GroundSeer(tm) touches upon. The following resources offer some more standard resources that provide a more authoritative view of the many issues related to AI, its safety and its interactions with humans.

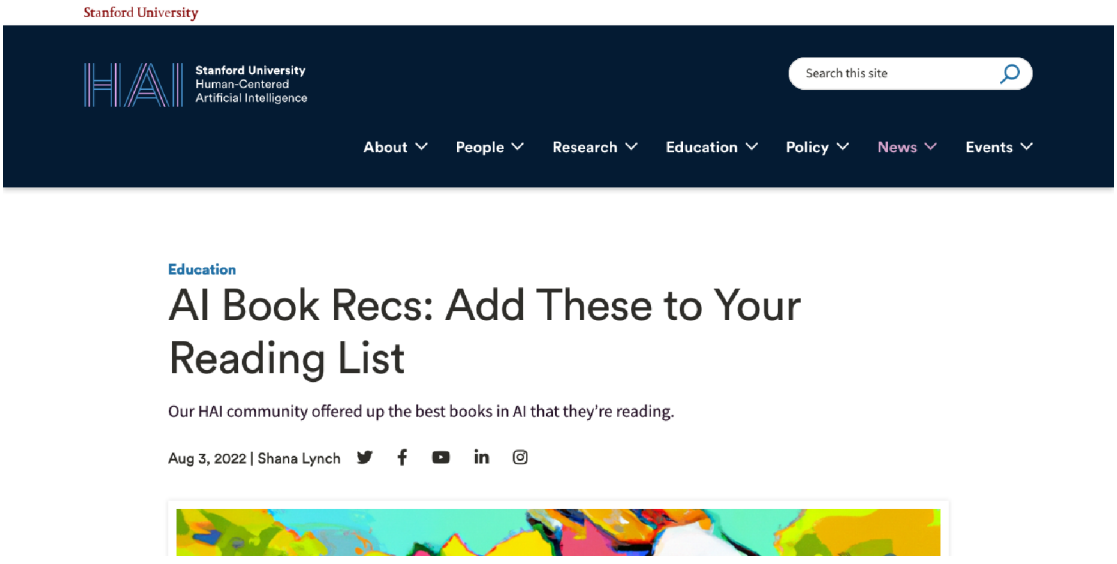
Stanford HAI Curated Summer Reading List

In[44]:=
```
haiURL = "https://hai.stanford.edu/news/ai-book-recs-add-these-your-reading-list"
```

Out[44]=
```
https://hai.stanford.edu/news/ai-book-recs-add-these-your-reading-list
```

In[45]:=
```
WebImage[haiURL, Method → "Firefox"]
```

Out[45]=



Understanding Algebraic Geometry

Cox, Little and O'Shea

Algebraic Statistics

AI Safety

Books

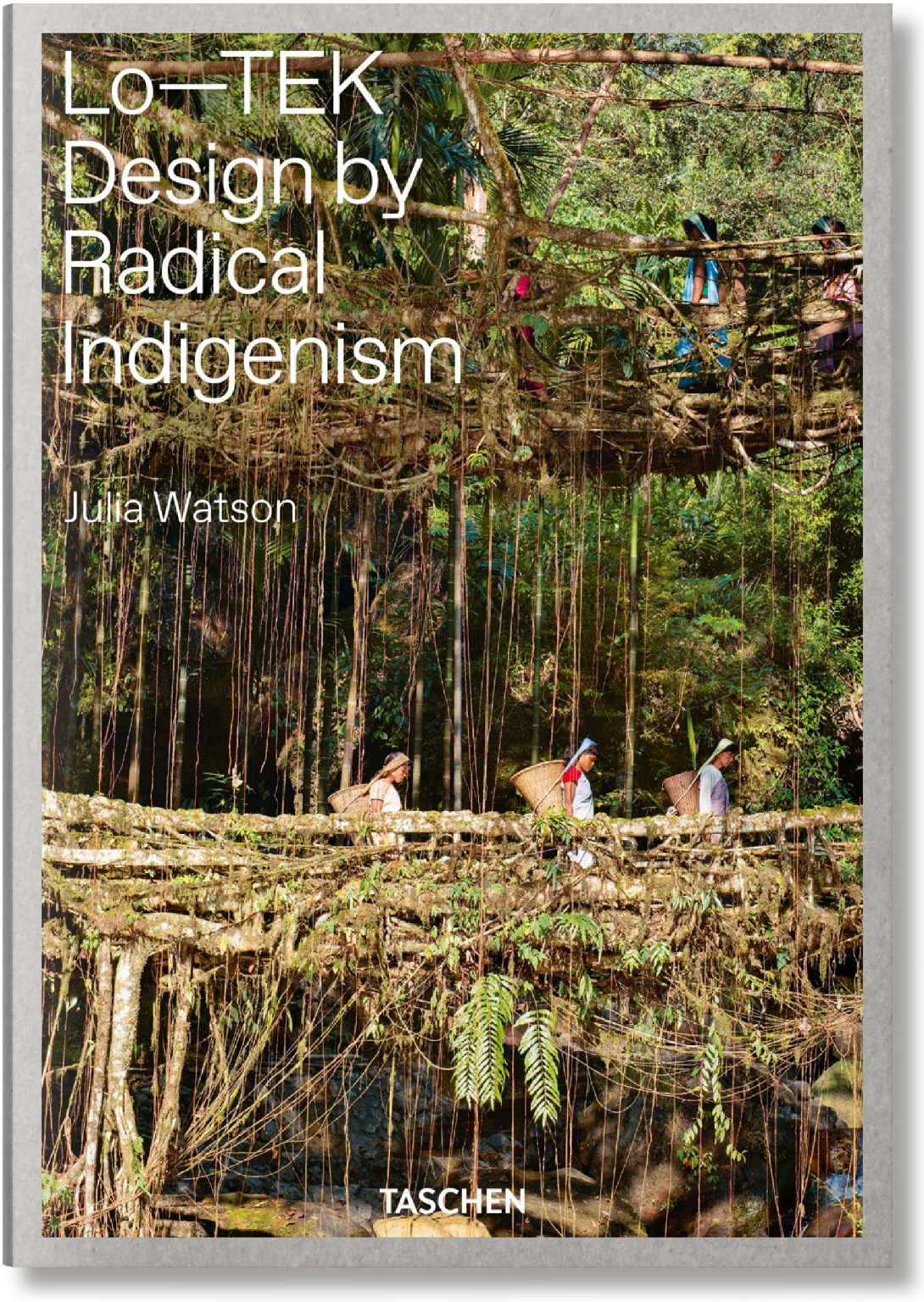
Institutes

Websites

Safe Design

Books

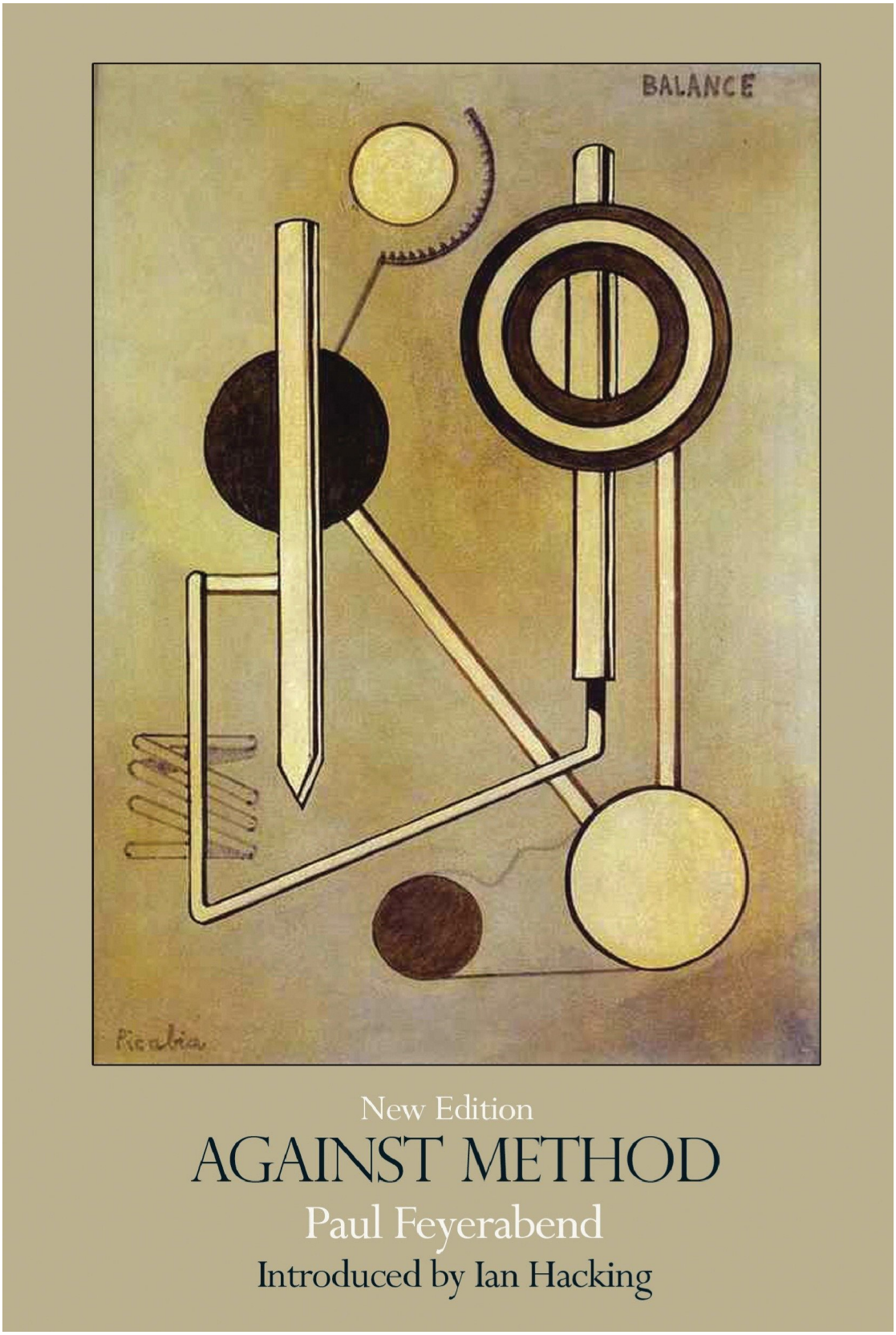
Lo-Tek: Design By Radical Indigenism by Julia Watson



Resilience is an important aspect of safe AI design. We must not think that only invention by AI researchers is going to be our only source of innovation. Technology and innovation are the birthright of humanity. All people across all time and space they inhabit have exhibited and exercised this trait

and continue to do so. This book takes the reader thru systems that have lasted for hundreds of years in some cases - like the bridge trees of Nepal pictured in its cover.

Against Method



This gadfly manifesto about how there is no established method to overcome the opinion of noisy judges in science contains the adage -

“There is no idea, however ancient and absurd, that is not capable of improving our knowledge. The whole history of thought is absorbed into science and is used for improving every single theory. Nor is political interference rejected. It may be needed to overcome the chauvinism of science that resists alternatives to the status quo. “

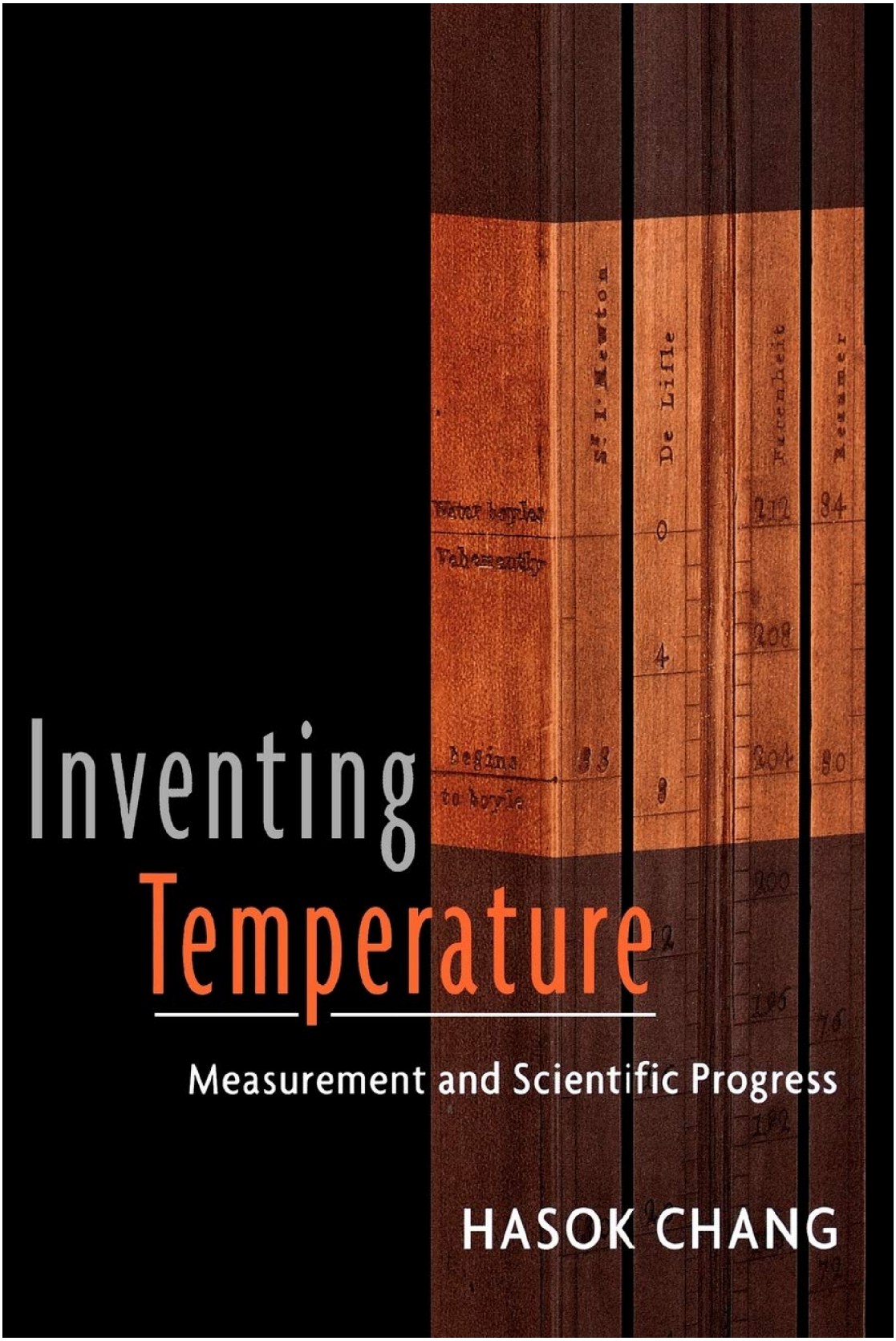
This submission follows this principle in various ways.

First, it provides a way to make all judges valuable for their self-assessment. Even wrong judges are okay as long as they are independent in their errors. Bad and middling classifiers can be used to evaluate stellar ones. This greatly increases the engineering uses of this idea.

Second, this very GitHub repository is a record of how new ideas take time and effort to materialize. This competition is the sort of thing Feyerabend would be proud to see. It allows a different channel for the entry of new ideas into the field of AI auditing and monitoring. We need more AI “Opens” like in tennis and golf.

Thirdly, encouraging a community of researchers and inventors in this area is a “political” act to some. To us it merely expresses socially the abstract mathematical core of our invention - we need many noisy judges of an unknown truth to understand it.

Inventing Temperature by Hasok Chang


Inventing
Temperature
Measurement and Scientific Progress
HASOK CHANG

The problem of the authority of purely empirical measurements to establish their own error was the concern of a now forgotten 19th century scientific luminary - Henri Regnault. His solution has relevance to the issue of AI safety and monitoring - purely empirical methods can be used to circumvent the need for further theory about the phenomena. The critique of Regnault's work was that he could not establish the precision of his measurements because to do so would require further theory about the phenomena being measured. His response to this was to establish an experimental protocol using four completely different thermometers, not just physically but also functioning with different mechanisms, expanding liquid, etc. His craftsmanship allowed him make them agree to 4 to 5 digits but then disagreed. This established - empirically - their error. No need to theorize about any models of the phenomena. The experimental attainment of high precision and then disagreement is all the authority needed. Any philosophical claims about absolute certainty is not possible are irrelevant to the safety engineer. Why not create more complex measurement protocols that can protect against increasingly complex failures? No error correcting code for bit errors is perfect, why should we expect error correcting codes for noisy judges to be different.