Getting to the truth

Author's remedies for the disinformation crisis will be controversial.

by **THOMAS LUMLEY**

DISTRUST: Big Data, Data-Torturing, and the Assault on Science, *by Gary Smith (Oxford University Press, $US30 hb)*

The theme of *Distrust* is attacks on both the process of science and its popular credibility. The book is generally accessible and readable, a polemic in the genre of Ben Goldacre's *Bad Science*. It will be most interesting to people without much knowledge of the area; many of the examples will already be familiar to those who follow discussions about the conduct of science.

Three reasons are given for a credibility crisis in science: disinformation, data torturing and data mining. Disinformation is part of a populist backlash and is mediated by the internet and social media and reinforced by those who profit from it. Data torturing is "driven by scientists' insistence on empirical evidence", and data mining is the automated version "fuelled by the big data and powerful computers that scientists created".

The stories of Gary Smith, an economics professor in California and author of *The AI Delusion*, with the possible exception of digressions on Bitcoin, are compelling illustrations of a problem for science.

One of the inside-cover blurbs compares the book to Carl Sagan's *The Demon-Haunted World*, a comparison that actually highlights one of the book's problems. *Distrust* has many historical examples, but its thesis of a modern crisis is somewhat weakened by the similarity of past and present errors.

The section on disinformation begins with a review of classical supernatural beliefs and conspiracy theories, as an introduction to the new post-fact, post-truth world. The internet is certainly a contributor to change. It's now easy for people to connect with like-minded folks outside their immediate social circle. There are flourishing internet communities of queer teens, blacksmiths, otaku, sourdough enthusiasts, Covid deniers and neo-Nazis.

As the inventors of the internet hoped, information circulates freely; disinformation does, too. During the height of the pandemic, then-prime minister Jacinda Ardern called for the government to be the "single source of truth" about Covid.

There was no real prospect of this happening, nor should it: individuals such as epidemiologist Professor Michael Baker and groups such as Te Rōpū Whakakaupapa Urutā, the National Māori Pandemic Group, were invaluable. The internet makes it harder to manufacture consent, and that is genuinely both a threat and an opportunity.

The remaining two main sections of the book, on data torturing and data mining, are basically statistical. Data torturing

The remedies start with raising the minimum age for social media to 16 or 18 and requiring identity certification.

has two themes: statistical significance without meaningful effects, and "fake" statistical significance from overly flexible research practices.

Data mining – automated analysis of large quantities of data – is not, I think, treated fairly in the book. There is a lot of argument that computers cannot understand or think, and therefore they cannot make good decisions or find meaningful associations. This is supported by the argument that big data analyses do very large numbers of comparisons and so must return many false positives. The conclusion does not follow, and I will give two examples where blind analysis of large databases has been genuinely useful.

There are three relatively new drugs, alirocumab, evolocumab and inclisiran, that reduce cholesterol more than older drugs. These drugs – not subsidised in New Zealand, although approved – inhibit a target, PCSK9, that was found by testing across the entire genomes of people with unusually high and low cholesterol. The algorithms used to discover PCSK9 didn't understand anything about cholesterol or heart attacks, but the results are real. Statistical control of false positives is entirely feasible in this sort of research. One might argue that looking for genetic correlations is a priori

sensible and so doesn't contravene the rule against data mining, but that makes the distinction between good and bad data mining quite fuzzy.

My second example is much more straightforward. Many, perhaps most, computational linguists would have argued 20 years ago that good machine translation could not be done just from correlations; some sort of understanding was intrinsically necessary. The likes of Google Translate, Deepl and Microsoft Translator have shown they were wrong.

The book ends with proposed remedies. Those for data mining and data torturing are reasonable, and many of them have been standard recommendations for some time, if not followed as much as they should be.

Smith's remedies for the disinformation crisis will be more controversial. They start with raising the minimum age for social media to 16 or 18 and requiring identity certification. The list goes on with a range of filtering and moderation proposals, such as "prohibit forwarding until a post has been reviewed and fact-checked".

The identity and age restrictions would reduce the value of social media enormously – even in the US, but especially in countries with less liberal speech laws – with the largest impact on those who have the fewest other ways to form communities.

The filtering restrictions, if feasible, would have benefits, but also come with important trade-offs.

During the recent brief Russian coup attempt, many people were bemoaning the loss of Twitter as a rapid source of information with clear provenance. Limits on amplification reduce active disinformation, but will hit the few who provide good information harder than the many who provide noise. ∎

Thomas Lumley is chair of biostatistics at the University of Auckland.
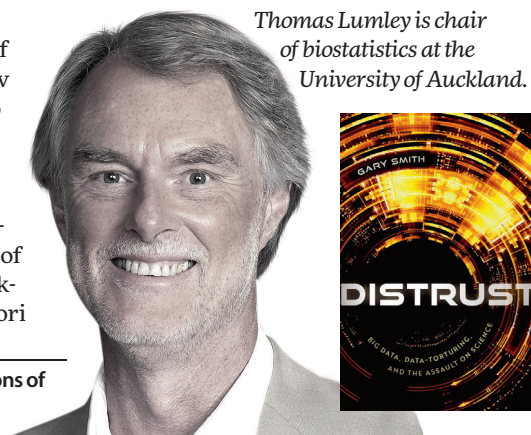
Gary Smith: offers compelling illustrations of a problem for science.

BESTSELLERS

MARGARET MEYER
THE WITCHING TIDE
THERE IS POWER IN SILENCE

TOP 5 FICTION

1. **ATLAS**, *by Lucinda Riley & Harry Whittaker (Macmillan)*
   Final story of the Seven Sisters

2. **KILLING MOON**, *by Jo Nesbo (Harvill Secker)*
   Harry Hole's back

3. **PET**, *by Catherine Chidgey (THWUP)*
   School-days psychological thriller

4. **CROSS DOWN**, *by James Patterson & Brendan Dubois (Century)*
   John Sampson is on his own

5. **THE FOURTH WING**, *by Rebecca Yarros (Piatkus)*
   Can Violet be a dragon rider?

TOP 5 NZ FICTION

1. **PET**, *by Catherine Chidgey (THWUP)*
   School-days psychological thriller

2. **THE AXEMAN'S CARNIVAL**, *by Catherine Chidgey (THWUP)*
   Human drama, with talking magpie

3. **BIRNAM WOOD**, *by Eleanor Catton (THWUP)*
   Billionaire's dastardly games

4. **THE WITCHING TIDE**, *by Margaret Meyer (Moa Press)*
   Witch-hunts in 17th-century East Anglia

5. **P.S. COME TO ITALY**, *by Nicky Pellegrino (Hachette)*
   Following your heart

(Bestsellers for week ending June 24)