Political scientists identify two components of trust: *information problems* and *commitment problems*. These two categories date back to the prisoner's dilemma, but more broadly can be applied to modeling agents in war (see bargaining theory), iterated games (see Axelrod's tournament), and protectionist economic policy.

While nearly all economists believe that free trade is (generally) a good thing, protectionist economic policy

For an example of the latter, the U.S.-China trade war is a good case study.

The importance of codes is readily apparent in history; see e.g. Alan Turing. Information theory provides a useful framework to enable trust in AI systems. There are three facets of "information-theoretic trust": *commpression*, *communication*, and *inference*. While these are standard topics in introductory texts like Thomas and Cover, it is worth expanding on the subtle connections between information theory and trust.

First, what does it take to safely encode a message? Shannon's source coding theorem states that N i.i.d. random variables with entropy $H(X)$ can be compressed into more than $NH(X)$ bits with negligible risk of information loss as $N \to \infty$. Conversely, if they are compressed into fewer than $NH(X)$ bits, it is virtually certain that information will be lost (MacKay). The language of compression is a powerful framework to analyze dimensionality reduction methods like PCA and autoencoders.

How much can one say over a noisy channel? Shannon's noisy-channel coding theorem states that there exists a non-negative number C (the channel capacity) with the following property. For any $\varepsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm such that the maximal probability of block error is $< \varepsilon$. This gives concrete

Finally, what priors should we use to model agents? There are many potential answers to this question, but information theory suggests that the "maximum-entropy distribution" makes the fewest assumptions. There are many expositions of this idea, see e.g. [...]. We will present

Suppose Alice is a researcher trying to understand some data. She is studying a phenomenon and wants to estimate a prior probability distribution among m possible outcomes. She has some prior information about what the phenomenon looks like.

To estimate the distribution, she runs the following experiment:

- Randomly distribute N quanta of probability, each worth $\frac{1}{N}$, among the m possibilities.
- Check if the probability assignment is consistent with her prior information. If inconsistent: reject and try again.
- If the assignment agrees with her prior information, her estimated prior distribution is given by

$$_i = \frac{n_i}{N}; \qquad i \in \{1, 2, \ldots, m\}$$

  where $_i$ is the probability of the i-th outcome, and n_i is the number of quanta that were assigned to the i-th proposition.

While model interpretability techniques (see e.g. X, Y, Z) have shed light on how to understand agents, information-theoretic models are still powerful tools we can apply to model the multi-agent case. And indeed, expositions

References

Cover, Thomas M., and Joy A. Thomas. *Elements of information theory.* John Wiley Sons, 2012.

Ganesh, Adithya C. "The principle of maximum entropy." *Stanford Mathematics Department, Directed reading program.*

MacKay, David JC, and David JC Mac Kay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Shannon, Claude Elwood. "A mathematical theory of communication." *Bell system technical journal* 27.3 (1948): 379-423.