Political scientists identify two categories of problems that impede trust: *information problems* and *commitment problems*. They are traditionally applied to the prisoner's dilemma, but they play a critical role in understanding problems of bargaining in war, iterated games, and economic policy.

For instance, the current U.S.-China trade war presents a setting in which two parties converge on ostensibly suboptimal policy (economic sanctions) because of these problems. First, each party lacks complete information on the long-term objectives of the other (e.g. geopolitical / economic agendas). Second, each party cannot be sure that the other will commit to policy changes (e.g. reduction in trade deficit, or improvement in Chinese market practices). Modelling such systems with information theory might play a role in building AI we can trust.

There are three core 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. [(cite my maxent article)]. The discrete uniform distribution is the maximum entropy prior with no further constraints. With fixed variance, the Gaussian is the maximum entropy prior; with fixed mean on $(0, \infty)$, the exponential distribution is the maximum entropy prior.

Suppose Alice is a researcher trying to understand some data. She is studying a phenomenon and wants to estimate a discrete prior over m possible outcomes, understand some constraints. To do so, 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.

The likelihood of any particular probability distribution is given by a multinomial coefficient, $\Pr(p) = \frac{N!}{n_1! n_2! \cdots n_m!} m^{-N}$. Using Stirling's approximation, it is not hard to show that the most likely prior will converge to the maximum entropy prior as $N \to \infty$ (the full derivation can be found in [...]).

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. Build a future with safe, trustworthy AI may require deep inquiry and understanding of the information-theoretic toolbox.

References

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

[2] Adithya C Ganesh. The principle of maximum entropy. *Stanford mathematics department: directed reading program*, 2019.

[3] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[4] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.