Political scientists identify two categories of problems that impede trust: *information problems* and *commitment problems* [5]. They are traditionally applied to the prisoner's dilemma, but they play a critical role in understanding problems of bargaining in war [8], iterated games [1], and economic policy [10].

For instance, the current U.S.–China trade war presents a setting in which two parties converge on an 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 the behavior of agents with information theory might play a role in building AI we can trust.

There are three core facets of "information-theoretic trust": *compression*, *communication*, and *inference*. While these are standard topics in introductory texts [3][9], it is worth expanding on the subtle connections between information theory and trust. I will elaborate on these three topics in the next few paragraphs.

First, compression plays a critical role in safe interaction with an ensemble of agents. Codes have played an essential role in history, e.g. see the prominent role of the Enigma machine in World War II [2]. Secure compression requires understanding how much we can safely encode in 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 $N \to \infty$ [11][9]. Conversely, if they are compressed into fewer than $NH(X)$ bits, it is virtually certain that information will be lost. The language of compression is a powerful framework to analyze dimensionality reduction models like principal component analysis [7] and autoencoders [4].

Second, safely interacting with AI requires trust in our communication channels. While the source coding theorem provides a useful framework to analyze the lossless case, in practice, we often deal with noisy channels. "Nuclear close calls" present useful case studies to examine in history, wherein nuclear weapons were "nearly" deployed by a legitimate authority [12]. In these settings, state actors are unsure of the intentions of others. The Norwegian rocket launch of 1995, described as a "poster child for nuclear dangers" (Tetrais 53 [**?**]), involved Norwegian and American scientists launching a rocket to study weather data. While information about this launch was sent to Moscow, before it reached the authorities, the launch was interpreted as a sign of a possible adversarial strike.

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 $< \varepsilon$. (...expand).

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.

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] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[2] Bernhelm Booß-Bavnbek and Jens Høyrup. *Mathematics and war*. Springer, 2003.

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

[4] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[5] Jeffry A Frieden, David A Lake, and Kenneth A Schultz. World politics. interests, interactions. *Institutions. WW Norton & Company, New York*, 2010.

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

[7] Bernhard C Geiger and Gernot Kubin. Relative information loss in the pca. In *2012 IEEE Information Theory Workshop*, pages 562–566. IEEE, 2012.

[8] David A Lake. Two cheers for bargaining theory: Assessing rationalist explanations of the iraq war. *International Security*, 35(3):7–52, 2010.

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

[10] Arvind Panagariya. International trade. *Foreign Policy*, pages 20–28, 2003.

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

[12] Bruno Tertrais. “on the brink”—really? revisiting nuclear close calls since 1945. *The Washington Quarterly*, 40(2):51–66, 2017.