**Tiago H. Falk**

**Lecture 2 – Trust in Human-Autonomous System Interaction: It Takes Two to Tango**

Abstract: In human-autonomous system interactions trust is an important aspect to be monitored. If trust levels are too low, humans will not rely on the assistive technology, making them underused and potentially increasing the risk of human errors. On the other side of the coin, if users over-trust a system, they can rely too much on automation, putting themselves at risk of missing certain threats or potentially propagating errors made by the system. As such, monitoring trust is crucial to ensure the right trade-off is met. Trust has been measured over the years using different modalities, including questionnaires or other quality-of-service indices. The IMPACTS trust model model, for example, shows the importance of seven elements to ensure humans remain trustworthy towards machines: intention (I), measurability (M), performance (P), adaptability (A), communication (C), transparency (T), and security (S). To date, monitoring trust has usually assumed that it is the user that needs to trust the system. In the era of deep learning based autonomous systems, however, trust should be a bidirectional entity, where two measures of trust are needed: one from the human towards the machine, and another for the machine towards the human commands. Indeed, deep learning based autonomous systems are known to be vulnerable to e.g., (i) adversarial attacks, (2) deepfakes, and more generally, (3) data from out-of-domain distributions. In this talk, I will highlight some of the signal processing tools we have developed over the years to help overcome these issues, as well as our most recent efforts at creating a bi-directional model of trust for human-autonomous system interactions.