Abstract:

One thing that distinguishes (theoretical) computer science from other scientific disciplines is its full-throated support of a fundamentally adversarial view of the universe. Malicious adversaries, with unbounded computational advantages, attempt to foil our algorithms at every turn and destroy their quantitative guarantees. However, there is one strange exception to this world view and it is this: the algorithm must accept its input as sacrosanct, and may never simply reject its input as illegitimate. But what if some inputs really are illegitimate? Is building a “bullshit detector” for algorithms a good idea?

To illustrate the power of the Bullshit Detection worldview, we give the first polynomial-time protocol for Byzantine Agreement that is resilient to \( f < \frac{n}{3} \) corruptions against an omniscient, computationally unbounded adversary in an asynchronous message-passing model. (This is the first improvement to Ben-Or and Bracha’s exponential time protocols from the 1980s that are resilient to \( f < \frac{n}{3} \) corruptions.) The key problem is to design a coin-flipping protocol in which corrupted parties (who chose the outcomes of their coins maliciously) are eventually detected via statistical tests.

We will also discuss other algorithmic contexts in which bullshit detection might be useful.