Machine learning (ML) can now decode what people think from brain scans. Recent studies have reconstructed continuous language from functional magnetic resonance imaging (fMRI), recovered visual imagery with over 90% accuracy, and read inner speech directly from motor cortex signals. These findings are published but so recent that their epistemological implications for psychiatry remain largely unexamined.

This matters because contemporary psychiatric epistemology holds that the mind is not reducible to the brain, it is embodied, embedded, and enacted through interaction with body and world. But ML decoders seem to bypass all of that, extracting mental content from neural signals alone, with no access to the patient's lived body, history, or social world. So why does it work? Does decoding success vindicate cerebrocentrism, suggesting that mental content is more brain-bound than embodied approaches assume, or are decoders merely picking up neural traces of fundamentally distributed processes?

This paper explores what this tension means for three issues: the epistemic status of prediction without understanding, the threat to first-person authority in psychiatric encounters, and the prospect that ML could reclassify mental disorders along biological lines that cut across our current diagnostic categories.