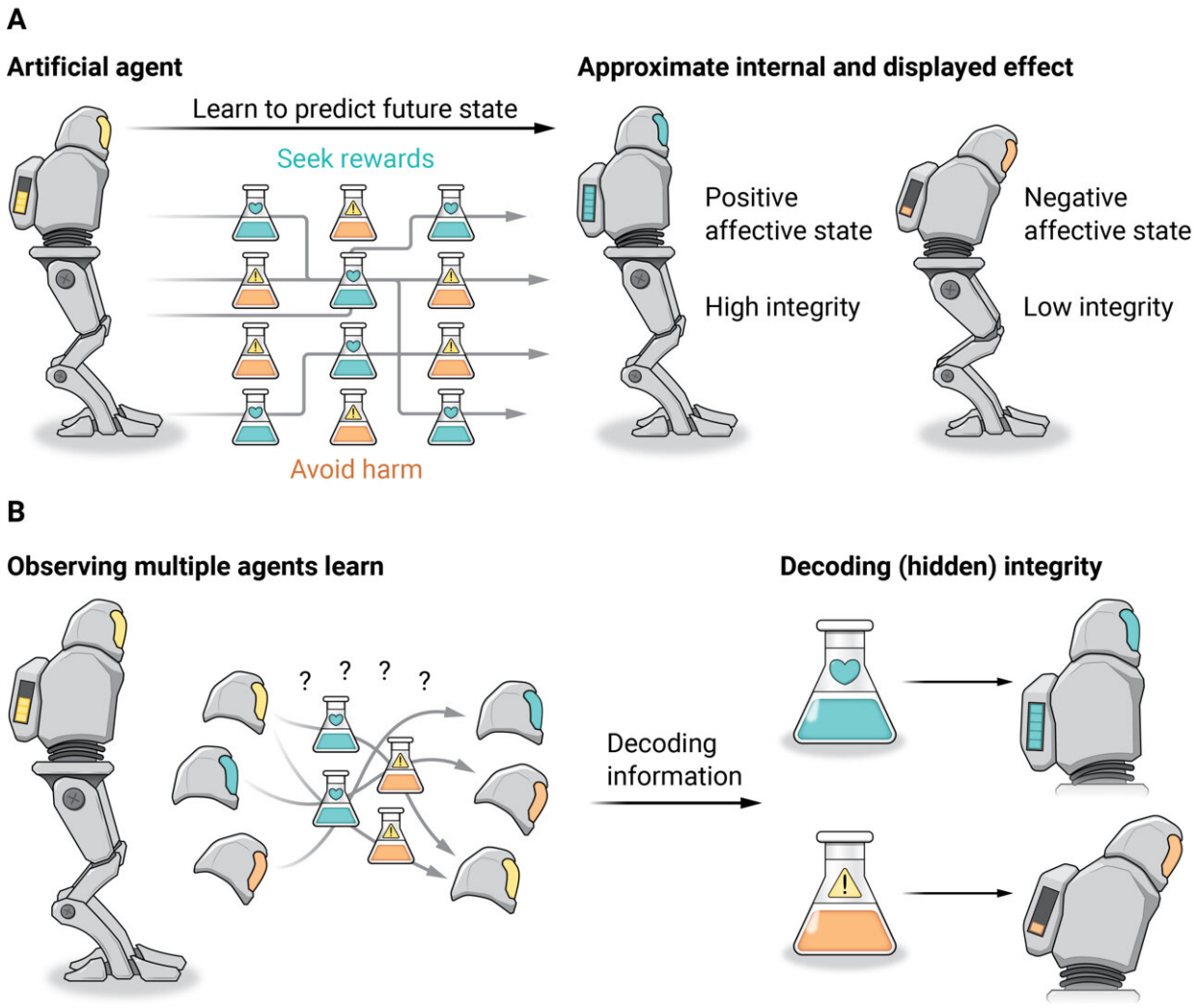
How to give AI-based robots empathy so they won't want to kill us

July 19 2023, by Bob Yirka



Developing artificial proxies for homeostasis, feeling and empathy. (A) The agent maintains its integrity within an environment by seeking rewards and avoiding harmful obstacles via predictive models of future states, and an

approximation of internal and displayed affect. (B) The agent must then leverage these models to decode and predict other agents' behavioral outcomes and internal affective states. Credit: Leonardo Christov-Moore and Nicco Reggente

A team of social scientists, neurologists and psychiatrists at the University of Southern California's Brain and Creativity Institute, working with colleagues from the Institute for Advanced Consciousness Studies, the University of Central Florida and the David Geffen School of Medicine at UCLA have published a Viewpoint piece in the journal *Science Robotics* outlining a new approach to giving robots empathy. In their paper, they suggest that traditional approaches may not work.

By nearly any measure, the introduction of ChatGPT and other AI apps like it has impacted modern society. They are being used for a broad range of purposes, but have instigated talk of curbing their development for fear that they may pose a threat to humans. To counter such arguments, some in the AI field have suggested that the means for preventing the development of such a scenario is simple—give the apps empathy. In this new paper, the authors agree with such an approach, but differ on how to mimic such an enigmatic human quality in a machine.

The current approach to conferring empathy to AI models centers on teaching them to see how humans behave under morally debatable scenarios and then to follow such behavior accordingly—and by hard-coding some rules into their machinery. But this approach, the authors argue, overlooks the role that self-preservation plays in human empathy. If a robot views video of a person experiencing a painful reaction to falling down, for example, it can be taught to mimic such a reaction as a way to connect with the person harmed, but it will be play-acting because it will not be feeling any empathy.

For that to happen, the robot would have to experience the kind of pain that can result from a fall. And that, the researchers suggest, is what must be done to get robots to understand why harming someone is bad, not coding a rule into their logic circuits. They are not suggesting that robots be programmed to feel real pain, though that might one day be an option, but instead to get them to see that their actions could have negative repercussions. They could have to face life without their human companion, for example, if they were to kill them. Or to be "killed" themselves because of what they have done. Doing so, they suggest, would involve giving robots the ability to suffer—an effective means of self-discipline if ever there was one.

More information: Leonardo Christov-Moore et al, Preventing antisocial robots: A pathway to artificial empathy, *Science Robotics* (2023). DOI: 10.1126/scirobotics.abq3658