DARPA Seeks to Make AI Robots More Moral and Less Potentially Psychopathic By Teaching Them Manners

Science (https://www.outerplaces.com/science) **Artificial Intelligence** (https://www.outerplaces.com/tag?tag=Artificial_Intelligence)

K.S. Anthony Monday, 28 January 2019 - 10:57AM

In the original 1979 *Alien,* the crew of the Nostromo is effectively sabotaged by its science officer, Ash, a Hyperdyne Systems 120-A/2 android programmed to ensure that Weyland-Yutani obtains a living xenomorph specimen, regardless of the cost to human life in doing so. In doing so, Ash deceives his human colleagues, violates security protocol (disobeying a direct order), and later tries to murder Ellen ***Ripley*** *(https://www.outerplaces.com/science-fiction/item/16412-alien-ellen-ripley-child-ridley-scott-prequel)* , thus leaving her with a profound distrust of androids and AI systems when she wakes up 57 years later having dispatched both Ash and the xenomorph. Unlike ***2001: A*** *(https://www.outerplaces.com/science-fiction/item/6484-bfis-re-release-trailer-for-2001-a-space-odyssey-revives-the-glory-of-sci-fi-gold)* ***Space*** *(https://www.outerplaces.com/tag?tag=Space)* Odyssey's Hal-9000 before it and the various ***Terminator*** *(https://www.outerplaces.com/tag?tag=Terminator)* models from the eponymous film series that followed, what makes Ash so astoundingly

compelling as an antagonist is because he's completely psychopathic in the way that only humans can be. Hal and the Terminator are recognizable as machines, but Ash doesn't show his milky innards until he is thoroughly dismantled and even then, he retains his brutal iciness. Because he looks human, we expect him to possess at least some of the "better angels of our nature," but he doesn't.

It's likely that the brainiacs at ***DARPA*** *(https://www.outerplaces.com/search?search=&q=DARPA)* have taken note of the critical eye cast on ***Artificial Intelligence*** *(https://www.outerplaces.com/tag?tag=Artificial Intelligence)* in popular culture. The agency is working on projects that make AI robots more human: better able to fit into society. One project in particular stands out.

In a ***2017 news release, DARPA*** *(https://www.darpa.mil/news-events/2017-05-31)* disclosed that it had contracted a group of researchers at Tufts and Brown University to decode how humans learn behavioral norms and act according to environmental and social contexts. The purpose of this project was to develop a method by which AI systems could be taught to assess behavioral norms, thus helping them better assimilate into human society and contexts. "The goal of this research effort was to understand and formalize human normative systems and how they guide human behavior," said DARPA program manager Reza Ghanadan, "so that we can set guidelines for how to design next-generation AI machines that are able to help and interact effectively with humans."

To address those needs, DARPA reported, "The team was able to create a cognitive-computational model of human norms in a representation that can be coded into machines, and developed a machine-learning algorithm that allows machines to learn norms in unfamiliar situations drawing on human data." In short, these researchers were able to create a basic formula for normative human behavior that works across various environmental and situational contexts. To better understand this, you have to consider the way an algorithm works. Despite the terror that the term inspires in social media managers, an algorithm is really just a set of rules or guidelines that describe how to perform a task: think of it as a huge flowchart filled with "If.... then...." statements.

In their ***published findings*** *(https://hrilab.tufts.edu/publications/sarathyetal2017cogsci.pdf)* , the researchers noted that there are rules that people intuit and generally abide by: certain behaviors that are prescribed or prohibited according to the situation, e.g. sitting in a public library and whispering, which also indicates a prohibition against talking loudly. In order to figure out how to teach an AI how to quickly assess these rules and act, the researchers developed expressions – formulas – to define norms, then used human participants to first generate norms in an array of contexts and then to detect norms across these various contexts. Finally, the researchers considered how people learn norms to begin with, taking into

account contextual clues that an AI might consider upon entering a new scenario. "Using a data representation format that incorporates several properties of human norm representation and learning," the team writes, "we then developed a novel algorithm for automatically learning context-sensitive norms from the human data." Whether or not DARPA has successfully implemented this algorithm into AI programming has yet to be seen.

It is worth noting that the team was led by ***Dr. Bertram Malle*** *(https://blog.cs.brown.edu/2017/06/06/darpa-looks-new-work-hcri-and-bertram-malle-teaching-robots-human-norms/)* of Brown University's Department of Cognitive, Linguistic, and Psychological Sciences. Besides a formidable body of research into cognitive and behavioral science, Dr. Malle's CV notes a particular interest in human-robot interactions, particularly "people's expectations of future robots; psychological mechanisms triggered by robot appearance (e.g., visual perspective taking); attempts to implement social-cognitive and moral competence in robots; conditions of optimal human-robot interaction (e.g., trust, explainability)."

As anthropomorphic technology becomes increasingly embedded in our lives – consider the fact that we no longer think anything of *talking* to proto-AIs like Siri and Alexa and we're quickly becoming used to the idea of humanoid robots – so do the inherent tensions. ***Forbes*** *(https://www.forbes.com/sites/jimvinoski/2018/09/03/will-a-robot-take-my-manufacturing-job-yes-no-and-maybe/#518f1250123f)* reported last year that robots have made humans all but obsolete for manufacturing jobs that once seemed like the last outpost for unskilled labor. Even iconoclasts like ***Elon Musk*** *(https://www.outerplaces.com/science/item/6564-elon-musk-creating-artificial-intelligence-is-akin-to-summoning-a-demon)* betray a certain phobia about AI that borders on superstition: in 2014, the ***SpaceX*** *(https://www.outerplaces.com/tag?tag=SpaceX)* and Tesla founder likened AI to "summoning the demon," which seems positively medieval in outlook, especially for a man who sent his car into orbit.

In a ***2015 editorial*** *(https://www.livescience.com/50349-how-to-raise-a-moral-robot.html)* , Dr. Malle discussed how we might develop moral robots and AI systems. In doing so, he illuminated some of the darkest corners of human behavior: the attitudes arising from fear that, as demonstrated time and time again, lead to actions that are illogical at their best and totally destructive at their worst. The real threat, Malle submits, is that robots will learn these fears from humans.

"Perhaps the greatest threat from robots comes from the greatest weakness of humans: hatred and conflict between groups. By and large, humans are cooperative and benevolent toward those whom they consider part of their group, but they can become malevolent and ruthless toward those outside their group. If robots learn such hostile sentiments and discriminatory actions, they may very well become a threat to humanity - or at least a threat to groups that the robot counts as 'outside' its community.

As ***recent experiments*** *(https://www.outerplaces.com/science/item/18551-mit-psychopath-ai-norman)* have shown, that threat is very real indeed. We would do well to take care that we instill these beings with the "better angels of our nature," lest Elon Musk's grim, medieval prophecy prove true.

(http://www.zergnet.com/i/3970824/62381/0/0/0/3)

5 Strange Places You Cannot Travel to Alone

(http://www.zergnet.com/i/3970824/62381/0/0/0/3)

(http://www.zergnet.com/i/3791938/62381/0/0/0/4)

The Little Girl from 'Mrs Doubtfire' is 31 Now and Gorgeous

(http://www.zergnet.com/i/3791938/62381/0/0/0/4)

(http://www.zergnet.com/i/3467042/62381/0/0/0/5)

Surprising Things Men Found Attractive 50 Years Ago

(http://www.zergnet.com/i/3467042/62381/0/0/0/5)

(http://www.zergnet.com/i/3476112/62381/0/0/0/6)

Why 'American Pickers' is a Total Sham

(http://www.zergnet.com/i/3476112/62381/0/0/0/6)

Ford's Smart Bed Prototype Forces Mattress Hogs to Stay in Their Lane

Science (https://www.outerplaces.com/science) **Artificial Intelligence** (https://www.outerplaces.com/tag?tag=Artificial_Intelligence)

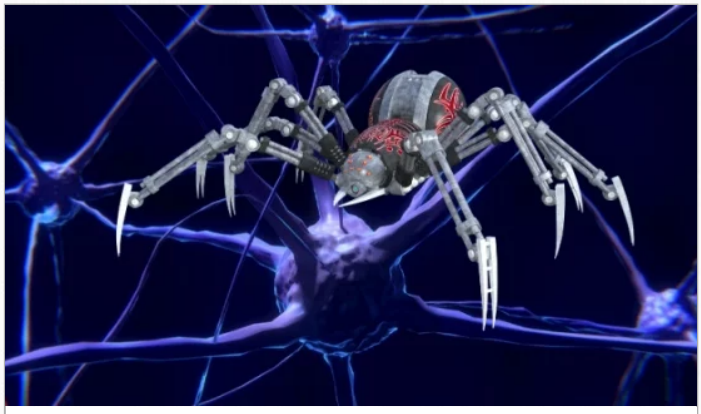
Ford's lane-keeping technology, installed in most of their new vehicles, detects road markings as you drive and gently nudges the steering wheel in the right direction if you should happen to go astray. What if you could apply that technology to other aspects of life? That's one of the questions that the auto manufacturer European branch is beginning to explore with its 'Ford Interventions' projects, which, according to ***Ford Europe*** *(https://fordeurope.blogspot.com/2019/02/fords-smart-bed-rolls-selfish-sleepers.html)* of Marketing Communications director Anthony Ireson, highlight some of the ways to "apply automotive expertise to tackle everyday – or in this case, every night – problems."

As the name suggests, the "Lane-Keeping Bed" solves the problem faced by sleepers who share a bed with someone who, unconsciously or not, takes up more than his or her share of the sleeping area, thus depriving their partner of a good night's rest. Keep in mind, this product is not for sale. It's simply a prototype created to demonstrate Ford's ingenuity and the myriad ways that technology can be adapted to improve our daily (and nightly) lives.

Related Stories

Poker-Playing AI Tapped For Military Use In $10M Pentagon Deal

(https://www.outerplaces.com/science/item/19258-ai-darpa-libratus-strategy-robot)

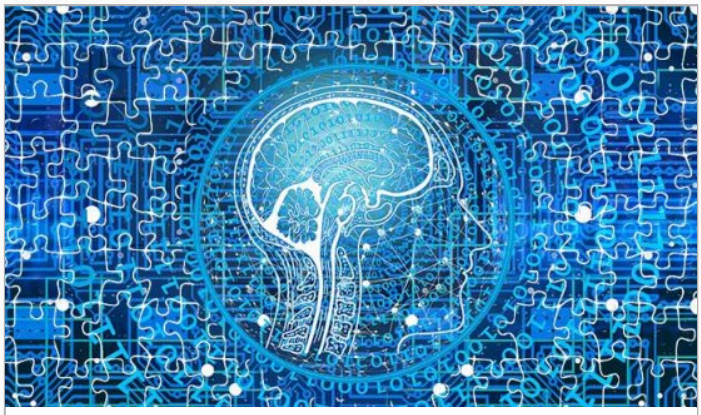


DARPA Wants To Use Insect Brains to Control Robots

(https://www.outerplaces.com/science/item/19240-darpa-insect-brains-robots)

DARPA Proposes An AI That Can Monitor The Entire World For Threats

(https://www.outerplaces.com/science/item/19232-darpa-kairos-artificial-intelligence)



DARPA Says the Biggest Obstacle to Effective Artificial Intelligence Is Common Sense

(https://www.outerplaces.com/science/item/18950-common-sense-artificial-intelligence)