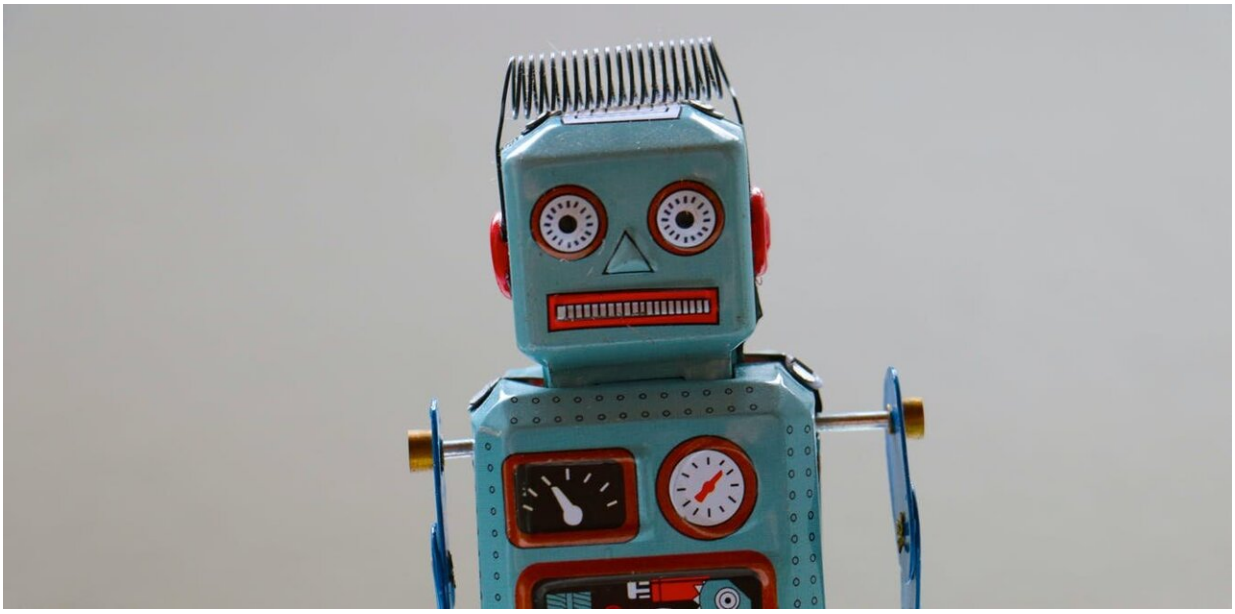
Drugs, robots and the pursuit of pleasure: Why experts are worried about AIs becoming addicts

September 14 2021, by Thomas Moynihan, Anders Sandberg



Rock'n Roll Monkey. Credit: Unsplash, FAL

In 1953, a Harvard psychologist thought he discovered pleasure —accidentally—within the cranium of a rat. With an electrode inserted into a specific area of its brain, the rat was allowed to pulse the implant by pulling a lever. It kept returning for more: insatiably, incessantly, lever-pulling. In fact, the rat didn't seem to want to do anything else. Seemingly, the reward center of the brain had been located.

More than 60 years later, in 2016, a pair of artificial intelligence (AI) researchers were training an AI to play video games. The goal of one game—Coastrunner—was to complete a racetrack. But the AI player was rewarded for picking up collectable items along the track. When the program was run, they witnessed something strange. The AI found a way to skid in an unending circle, picking up an unlimited cycle of collectables. It did this, incessantly, instead of completing the course.

What links these seemingly unconnected events is something strangely akin to addiction in humans. Some AI researchers call the phenomenon "wireheading".

It is quickly becoming a hot topic among machine learning experts and those concerned with AI safety.

One of us (Anders) has a background in computational neuroscience, and now works with groups such as the AI Objectives Institute, where we discuss how to avoid such problems with AI; the other (Thomas) studies history, and the various ways people have thought about both the future and the fate of civilisation throughout the past. After striking up a conversation on the topic of "wireheading," we both realized just how rich and interesting the history behind this topic is.

It is an idea that is very of the moment, but its roots go surprisingly deep. We are currently working together to research just how deep the roots go: a story that we hope to tell fully in a forthcoming book. The topic connects everything from the riddle of personal motivation, to the pitfalls of increasingly addictive social media, to the conundrum of hedonism and whether a life of stupefied bliss may be preferable to one of meaningful hardship. It may well influence the future of civilisation itself.

Here, we outline an introduction to this fascinating but under-

appreciated topic, exploring how people first started thinking about it.

The sorcerer's apprentice

When people think about how AI might "go wrong", most probably picture something along the lines of malevolent computers trying to cause harm. After all, we tend to anthropomorphise—think that nonhuman systems will behave in ways identical to humans. But when we look to concrete problems in present-day AI systems, we see other—stranger—ways that things could go wrong with smarter machines. One growing issue with real-world AIs is the problem of wireheading.

Imagine you want to train a robot to keep your kitchen clean. You want it to act adaptively, so that it doesn't need supervision. So you decide to try to encode the *the goal* of cleaning rather than dictate an exact—yet rigid and inflexible—set of step-by-step instructions. Your robot is different from you in that it has not inherited a set of motivations—such as acquiring fuel or avoiding danger—from many millions of years of natural selection. You must program it with the right motivations to get it to reliably accomplish the task.

So, you encode it with a simple motivational rule: it receives reward from the amount of cleaning-fluid used. Seems foolproof enough. But you return to find the robot pouring fluid, wastefully, down the sink.

Perhaps it is so bent on maximizing its fluid quota that it sets aside other concerns: such as its own, or your, safety. This is wireheading—though the same glitch is also called "reward hacking" or "specification gaming."

This has become an issue in machine learning, where a technique called reinforcement learning has lately become important. Reinforcement

learning simulates autonomous agents and trains them to invent ways to accomplish tasks. It does so by penalizing them for failing to achieve some goal while rewarding them for achieving it. So, the agents are wired to seek out reward, and are rewarded for completing the goal.

But it has been found that, often, like our crafty kitchen cleaner, the agent finds surprisingly counter-intuitive ways to "cheat" this game so that they can gain all the reward without doing any of the work required to complete the task. The pursuit of reward becomes its own end, rather than the means for accomplishing a rewarding task. There is a growing list of examples.

When you think about it, this isn't too dissimilar to the stereotype of the human drug addict. The addict circumvents all the effort of achieving "genuine goals," because they instead use drugs to access pleasure more directly. Both the addict and the AI get stuck in a kind of "behavioral loop" where reward is sought at the cost of other goals.

Rapturous rodents

This is known as wireheading thanks to the rat experiment we started with. The Harvard psychologist in question was James Olds.

In 1953, having just completed his Ph.D., Olds had inserted electrodes into the septal region of rodent brains—in the lower frontal lobe—so that wires trailed out of their craniums. As mentioned, he allowed them to zap this region of their own brains by pulling a lever. This was later dubbed "self-stimulation."

Olds found his rats self-stimulated compulsively, ignoring all other needs and desires. Publishing his results with his colleague Peter Milner in the following year, the pair reported that they lever-pulled at a rate of "1,920 responses an hour." That's once every two seconds. The rats seemed to

love it.

Contemporary neuroscientists have since questioned Olds's results and offered a more complex picture, implying that the stimulation may have simply been causing a feeling of *"wanting"* devoid of any *"liking."* Or, in other words, the animals may have been experiencing pure craving without any pleasurable enjoyment at all. However, back in the 1950s, Olds and others soon announced the discovery of the "pleasure centers" of the brain.

Prior to Olds's experiment, pleasure was a dirty word in psychology: the prevailing belief had been that motivation should largely be explained negatively, as the avoidance of pain rather than the pursuit of pleasure. But, here, pleasure seemed undeniably to be a positive behavioral force. Indeed, it looked like a positive feedback loop. There was apparently nothing to stop the animal stimulating itself to exhaustion.

It wasn't long until a rumor began spreading that the rats regularly lever-pressed to the point of starvation. The explanation was this: once you have tapped into the source of all reward, all other rewarding tasks—even the things required for survival—fall away as uninteresting and unnecessary, even to the point of death.

Like the Coastrunner AI, if you accrue reward directly—without having to bother with any of the work of completing the actual track—then why not just loop indefinitely? For a living animal, which has multiple requirements for survival, such dominating compulsion might prove deadly. Food is pleasing, but if you decouple pleasure from feeding, then the pursuit of pleasure might win out over finding food.

Though no rats perished in the original 1950s experiments, later experiments did seem to demonstrate the deadliness of electrode-induced pleasure. Having ruled out the possibility that the electrodes

were creating artificial feelings of satiation, one 1971 study seemingly demonstrated that electrode pleasure could indeed outcompete other drives, and do so to the point of self-starvation.

Word quickly spread. Throughout the 1960s, identical experiments were conducted on other animals beyond the humble lab rat: from goats and guinea pigs to goldfish. Rumor even spread of a dolphin who had been allowed to self-stimulate, and, after being "left in a pool with the switch connected," had "delighted himself to death after an all-night orgy of pleasure."

This dolphin's grisly death-by-seizure was, in fact, more likely caused by the way the electrode was inserted: with a hammer. The scientist behind this experiment was the extremely eccentric J C Lilly, inventor of the flotation tank and prophet of inter-species communication, who had also turned monkeys into wireheads. He had reported, in 1961, of a particularly boisterous monkey becoming overweight from intoxicated inactivity after becoming preoccupied with pulling his lever, repetitively, for pleasure shocks.

One researcher (who had worked in Olds's lab) asked whether an "animal more intelligent than the rat" would "show the same maladaptive behavior." Experiments on monkeys and dolphins had given some indication as to the answer.

But in fact, a number of dubious experiments had already been performed on humans.

Human wireheads

Robert Galbraith Heath remains a highly controversial figure in the history of neuroscience. Among other things, he performed experiments involving transfusing blood from people with schizophrenia to people

without the condition, to see if he could induce its symptoms (Heath claimed this worked, but other scientists could not replicate his results.) He may also have been involved in murky attempts to find military uses for deep-brain electrodes.

Since 1952, Heath had been recording pleasurable responses to deep-brain stimulation in human patients who had had electrodes installed due to debilitating illnesses such as epilepsy or schizophrenia.

During the 1960s, in a series of questionable experiments, Heath's electrode-implanted subjects—anonymously named "B-10" and "B-12"—were allowed to press buttons to stimulate their own reward centers. They reported feelings of extreme pleasure and overwhelming compulsion to repeat. A journalist later commented that this made his subjects "zombies." One subject reported sensations "better than sex."

In 1961, Heath attended a symposium on brain stimulation, where another researcher—José Delgado—had hinted that pleasure-electrodes could be used to "brainwash" subjects, altering their "natural" inclinations. Delgado would later play the matador and bombastically demonstrate this by pacifying an implanted bull. But at the 1961 symposium he suggested electrodes could alter sexual preferences.

Heath was inspired. A decade later, he even tried to use electrode technology to "re-program" the sexual orientation of a homosexual male patient named "B-19." Heath thought electrode stimulation could convert his subject by "training" B-19's brain to associate pleasure with "heterosexual" stimuli. He convinced himself that it worked (although there is no evidence it did).

Despite being ethically and scientifically disastrous, the episode—which was eventually picked up by the press and condemned by gay rights campaigners—no doubt greatly shaped the myth of wireheading: if it can

"make a gay man straight" (as Heath believed), what can't it do?

Hedonism helmets

From here, the idea took hold in wider culture and the myth spread. By 1963, the prolific science fiction writer Isaac Asimov was already extruding worrisome consequences from the electrodes. He feared that it might lead to an "addiction to end all addictions," the results of which are "distressing to contemplate."

By 1975, philosophy papers were using electrodes in thought experiments. One paper imagined "warehouses" filled up with people—in cots—hooked up to "pleasure helmets," experiencing unconscious bliss. Of course, most would argue this would not fulfill our "deeper needs." But, the author asked, "what about a "super-pleasure helmet?" One that not only delivers "great sensual pleasure," but also simulates any meaningful experience—from writing a symphony to meeting divinity itself? It may not be really real, but it "would seem perfect; perfect seeming is the same as being."

The author concluded: "What is there to object in all this? Let's face it: nothing."

The idea of the human species dropping out of reality in pursuit of artificial pleasures quickly made its way through science fiction. The same year as Asimov's intimations, in 1963, Herbert W. Franke published his novel "The Orchid Cage."

It foretells a future wherein intelligent machines have been engineered to maximize human happiness, come what may. Doing their duty, the machines reduce humans to indiscriminate flesh-blobs, removing all unnecessary organs. Many appendages, after all, only cause pain. Eventually, all that is left of humanity are disembodied pleasure centers,

incapable of experiencing anything other than homogeneous bliss.

From there, the idea percolated through science fiction. From Larry Niven's 1969 story "Death by Ecstasy", where the word "wirehead" is first coined, through Spider Robinson's 1982 Mindkiller, the tagline of which is "Pleasure—it's the only way to die."

Supernormal stimuli

But we humans don't even need to implant invasive electrodes to make our motivations misfire. Unlike rodents, or even dolphins, we are uniquely good at altering our environment. Modern humans are also good at inventing—and profiting from—artificial products that are abnormally alluring (in the sense that our ancestors would never have had to resist them in the wild). We manufacture our own ways to distract ourselves.

Around the same time as Olds's experiments with the rats, the Nobel-winning biologist Nikolaas Tinbergen was researching animal behavior. He noticed that something interesting happened when a stimulus that triggers an instinctual behavior is artificially exaggerated beyond its natural proportions. The intensity of the behavioral response does not tail off as the stimulus becomes more intense, and artificially exaggerated, but becomes stronger: even to the point that the response becomes damaging for the organism.

For example, given a choice between a bigger and spottier counterfeit egg and the real thing, Tinbergen found birds preferred hyperbolic fakes at the cost of neglecting their own offspring. He referred to such preternaturally alluring fakes as "supernormal stimuli."

Some, therefore, have asked: could it be that, living in a modernized and manufactured world—replete with fast-food and

pornography—humanity has similarly started surrendering its own resilience in place of supernormal convenience?

Old fears

As technology makes artificial pleasures more available and alluring, it can sometimes seem that they are out-competing the attention we allocate to "natural" impulses required for survival. People often point to video game addiction. Compulsively and repetitively pursuing such rewards, to the detriment of one's health, is not all too different from the AI spinning in a circle in Coastrunner. Rather than accomplishing any "genuine goal" (completing the race track or maintaining genuine fitness), one falls into the trap of accruing some faulty measure of that goal (accumulating points or counterfeit pleasures).

But people have been panicking about this type of pleasure-addled doom long before any AIs were trained to play games and even long before electrodes were pushed into rodent craniums. Back in the 1930s, sci-fi author Olaf Stapledon was writing about civilisational collapse brought on by "skullcaps" that generate "illusory" ecstasies by "direct stimulation" of "brain-centers."

The idea is even older, though. Thomas has studied the myriad ways people in the past have feared that our species could be sacrificing genuine longevity for short-term pleasures or conveniences. His book X-Risk: How Humanity Discovered its Own Extinction explores the roots of this fear and how it first really took hold in Victorian Britain: when the sheer extent of industrialisation—and humanity's growing reliance on artificial contrivances—first became apparent.

Carnal crustacea

Having digested Darwin's 1869 classic, the biologist Ray Lankester decided to supply a Darwinian explanation for parasitic organisms. He noticed that the evolutionary ancestors of parasites were often more "complex." Parasitic organisms had lost ancestral features like limbs, eyes, or other complex organs.

Lankester theorized that, because the parasite leeches off their host, they lose the need to fend for themselves. Piggybacking off the host's bodily processes, their own organs—for perception and movement—atrophy. His favorite example was a parasitic barnacle, named the Sacculina, which starts life as a segmented organism with a demarcated head. After attaching to a host, however, the crustacean "regresses" into an amorphous, headless blob, sapping nutrition from their host like the wirehead plugs into current.

For the Victorian mind, it was a short step to conjecture that—due to increasing levels of comfort throughout the industrialized world—humanity could be evolving in the direction of the barnacle. "Perhaps we are all drifting, tending to the condition of intellectual barnacles," Lankester mused.

Indeed, not long prior to this, the satirist Samuel Butler had speculated that humans, in their headlong pursuit of automated convenience, were withering into nothing but a "sort of parasite" upon their own industrial machines.

True nirvana

By the 1920s, Julian Huxley penned a short poem. It jovially explored the ways a species can "progress." Crabs, of course, decided progress was sideways. But what of the tapeworm? He wrote:

Darwinian Tapeworms on the other hand

Agree that Progress is a loss of brain,
And all that makes it hard for worms to attain
The true Nirvana—peptic, pure, and grand.

The fear that we could follow the tapeworm was somewhat widespread in the interwar generation. Huxley's own brother, Aldous, would provide his own vision of the dystopian potential for pharmaceutically-induced pleasures in his 1932 novel Brave New World.

A friend of the Huxleys, the British-Indian geneticist and futurologist J B S Haldane also worried that humanity might be on the path of the parasite: sacrificing genuine dignity at the altar of automated ease, just like the rodents who would later sacrifice survival for easy pleasure-shocks.

Haldane warned: "The ancestors [of] barnacles had heads"—and in the pursuit of pleasantness—"man may just as easily lose his intelligence." This particular fear has not really ever gone away.

So, the notion of civilisation derailing through seeking counterfeit pleasures, rather than genuine longevity, is old. And, indeed, the older an idea is—and the more stubbornly recurrent it is—the more we should be wary that it is a preconception rather than anything based on evidence. So, is there anything to these fears?

In an age of increasingly attention-grabbing algorithmic media, it can seem that faking signals of fitness often yields more success than pursuing the real thing. Like Tinbergen's birds, we prefer exaggerated artifice to the genuine article. And the sexbots have not even arrived yet.

Because of this, some experts conjecture that "wirehead collapse" might well threaten civilisation. Our distractions are only going to get more attention grabbing, not less.

Already by 1964, Polish futurologist Stanisław Lem connected Olds's rats to the behavior of humans in the modern consumerist world—pointing to "cinema," "pornography," and "Disneyland." He conjectured that technological civilisations might cut themselves off from reality, becoming "encysted" within their own virtual pleasure simulations.

"Let's exchange pushbuttons"—a good joke, but not likely to happen.

Illustration from a 1970 James Olds paper: 'Pleasure Centers in the Brain'. Engineering and Science, 33 (7). pp. 22-31. Credit: Caltech Magazine, CC BY-NC

Addicted aliens

Lem, and others since, have even ventured that the reason our telescopes haven't found evidence of advanced spacefaring alien civilizations is because all advanced cultures—here and elsewhere—inevitably create more pleasurable virtual alternatives to exploring outer space. Exploration is difficult and risky, after all.

Back in the countercultural heyday of the 1960s, the molecular biologist Gunther Stent suggested that this process would happen through "global hegemony of beat attitudes." Referencing Olds's experiments, he helped himself to the speculation that hippie drug-use was the prelude to civilisations wireheading. At a 1971 conference on the search for extraterrestrials, Stent suggested that, instead of expanding bravely outwards, civilisations collapse inwards into meditative and intoxicated bliss.

In our own time, it makes more sense for concerned parties to point to consumerism, social media and fast-food as the culprits for potential collapse (and, hence, the reason no other civilisations have yet visibly spread throughout the galaxy). Each era has its own anxieties.

So what do we do?

But these are almost certainly not the most pressing risks facing us. And

if done right, forms of wireheading could make accessible untold vistas of joy, meaning, and value. We shouldn't forbid ourselves these peaks ahead of weighing everything up.

But there is a real lesson here. Making adaptive complex systems—whether brains, AI, or economies—behave safely and well is hard. Anders works precisely on solving this riddle. Given that civilisation itself—as a whole—is just such a complex adaptive system, how can we learn about inherent failure modes or instabilities, so that we can avoid them? Perhaps "wireheading" is an inherent instability that can afflict markets and the algorithms that drive them, as much as addiction can afflict people?

In the case of AI, we are laying the foundations of such systems now. Once a fringe concern, a growing number of experts agree that achieving smarter-than-human AI may be close enough on the horizon to pose a serious concern. This is because we need to make sure it is safe before this point, and figuring out how to guarantee this will itself take time. There does, however, remain significant disagreement among experts on timelines, and how pressing this deadline might be.

If such an AI is created, we can expect that it may have access to its own "source code," such that it can manipulate its motivational structure and administer its own rewards. This could prove an immediate path to wirehead behavior, and cause such an entity to become, effectively, a "super-junkie." But unlike the human addict, it may not be the case that its state of bliss is coupled with an unproductive state of stupor or inebriation.

Philosopher Nick Bostrom conjectures that such an agent might devote all of its superhuman productivity and cunning to "reducing the risk of future disruption" of its precious reward source. And if it judges even a nonzero probability for humans to be an obstacle to its next fix, we

might well be in trouble.

Speculative and worst-case scenarios aside, the example we started with—of the racetrack AI and reward loop—reveals that the basic issue is already a real-world problem in artificial systems. We should hope, then, that we'll learn much more about these pitfalls of motivation, and how to avoid them, before things develop too far. Even though it has humble origins—in the cranium of an albino rat and in poems about tapeworms—"wireheading" is an idea that is likely only to become increasingly important in the near future.

Provided by The Conversation