NAUTILUS

CULTURE | SOCIOLOGY

# Is Tribalism a Natural Malfunction?

*What computers teach us about getting along.*

BY SIMON DEDEO
ILLUSTRATION BY FRANCESCO IZZO
SEPTEMBER 14, 2017

ADD A COMMENT FACEBOOK TWITTER EMAIL SHARING REDDIT STUMBLEUPON TUMBLR POCKET

From an office at Carnegie Mellon, my colleague John Miller and I had evolved a computer program with a taste for genocide.

This was certainly not our intent. We were not scholars of race, or war. We were interested in the emergence of primitive cooperation. So we built machines that lived in an imaginary society, and made them play a game with each other—one known to engender complex social behavior just as surely as a mushy banana makes fruit flies.

The game is called Prisoner's Dilemma. It takes many guises, but it is at heart a story about two individuals that can choose to cooperate or to cheat. If they both cheat, they both suffer. If they both cooperate, they both prosper. But if one tries to cooperate while the other cheats, the cheater prospers even more.

The game has a generality that appeals to a political philosopher, but a rigorous specificity that makes it possible to guide computer simulations. As a tool for the mathematical study of human behavior, it is the equivalent of Galileo's inclined plane, or Gregor Mendel's pea plants. Do you join the strike, or sneak across the picket line? Rein in production to keep prices high, or undercut the cartel and flood the market? Pull your weight in a study group, or leave the work to others?

*Woe to those who did not know the code.*

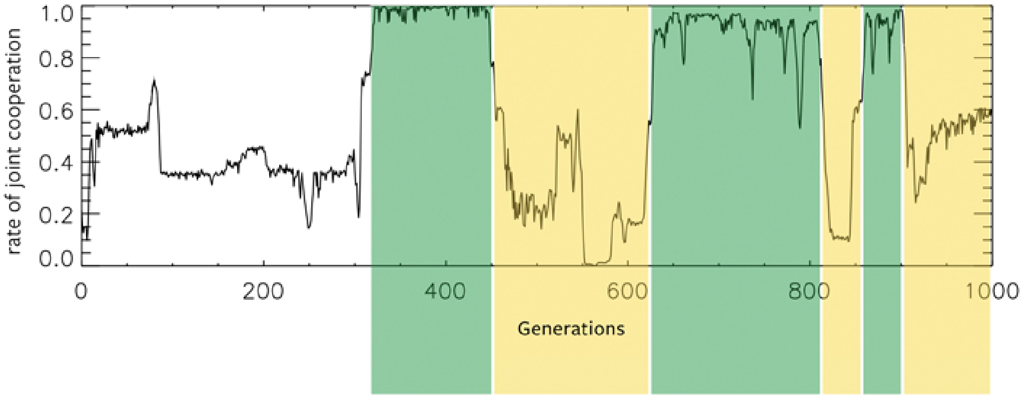
Our simulation was simple: In a virtual world, decision-making machines with limited powers of reasoning played the game over and over. We, as the unforgiving account-keepers, rewarded the ones who prospered and punished the ones who did not. Successful machines passed their strategies to the next generation, with the occasional slight variations designed to imitate the blind distortions typical of cultural evolution.

We also gave the machines a simple language to think with and enough resources to have memories and to act on them. Each generation, paired machines faced each other multiple times. This is how life appears to us: We encounter our trading partners over and over, and how we treat them has consequences. Our model for the world was two Robinson Crusoes encountering each other on the sands.

When we ran these little societies forward, we expected to confirm what many believed to the optimal strategy for playing Prisoner's Dilemma: tit-for-tat. A machine playing this strategy begins by keeping its promises, but retaliates against an instance of cheating by cheating, once, in return. Tit-for-tat is the playground rule of honor: Treat others well, unless they give you reason otherwise—and be reasonably quick to forgive.

Yet when we looked at the output of our simulations, where the strategies were free to evolve in arbitrary directions, we saw something very different. After an early, chaotic period, a single machine would rise rapidly to dominance, taking over its imaginary world for hundreds of generations until, just as suddenly, it collapsed, sending the world into a chaos of conflict out of which the next cycle arose. An archaeologist of such a world would have encountered thick layers of prosperity alternating with eras of ash and bone.

Instead of an orderly playground ruled by cautious, prideful cooperators, the population produced bizarre configurations that made no sense to us. That is, until one evening, in the office and after filling up pads of graph paper, we stumbled onto the truth. The dominant machines had taken players' actions to be a code by which they could recognize when they were faced with copies of themselves.



**SHIBBOLETH MACHINES:** Simulations of our machines show initial levels of apparently random behavior giving way, around generation 300, to high rates of cooperation that coincide with near-complete domination by a single machine that drives others to extinction. This enforced cooperation collapses around generation 450. From then on, the system alternates between these two extremes. Green and yellow bands correspond to eras of high and low cooperation, respectively.

In the opening moves of the game, they would tap out a distinct pattern: cooperate, cheat, cheat, cooperate, cheat, cooperate (for example). If their opponent responded in exactly the same fashion, cheating when they cheated, cooperating when they cooperated, they would eventually switch to a phase of permanent cooperation, rewarding the opponent with the benefits of action to mutual advantage.

Woe, however, to those who did not know the code. Any deviation from the expected sequence was rewarded with total and permanent war. Such a response might take both machines down, in a kind of a digital suicide attack. Because the sequence was so hard to hit upon by accident, only the descendants of ruling machines could profit from the post-code era of selfless cooperation. All others were killed off, including those using the tit-for-tat strategy. This domination would last until enough errors accumulated in the code handed down between generations for dominant machines to stop recognizing each other. Then, they would turn against each other as viciously as they once turned against outsiders, in a kind of population-level autoimmune disease.

As long as the codes lasted we called them Shibboleths, after the tribal genocide recounted in the Old Testament Book of Judges:

> And the Gileadites took the passages of Jordan before the Ephraimites: and it was *so*, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, *Art* thou an Ephraimite? If he said, Nay; / Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce *it* right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand.

Shibboleths are a common feature of human culture and conflict. Finns who could not pronounce *yksi* (meaning "one") were identified as Russians during the Finnish Civil War. Tourists in downtown Manhattan quickly out themselves if they pronounce Houston Street like the city in Texas.

Here our machines had used them to dominate a population so effectively that no others could survive. Even after the era was over, it was their descendants that inherited the ashes. The blind hand of evolution had found a simple, if vicious, solution.

It was a stark and brutal social landscape. But we had given our machines very limited resources to think with. How would two perfectly rational machines act in a conflict, if they each knew the other was similarly perfectly rational? By the very nature of rationality, two completely rational beings, confronted with the same problem, ought to behave in the same fashion. Knowing this, each would choose to cooperate—but not out of altruism. Each would recognize that if it were to cheat, its opponent would too, making them both losers in the game.

The two endpoints establish a spectrum. At one end are our minimally-calculating machines, parochial zero-points of culture that naturally, we found, distilled down to a vicious tribalism. At the other end is the inevitable cooperation of the perfectly rational agent.

On this line between beastly machines and angelic rationality, where do we find the human species?

If we humans are super-rational, or at least on our way there, there is reason to be optimistic. Francis Fukuyama might have been thinking along these lines when he penned his end-of-history thesis in 1992. Though Fukuyama's argument was rooted in 19th-century German philosophers such as Friedrich Nietzsche and Georg Wilhelm Friedrich Hegel, we might rewrite it this way: A sufficiently complex simulation of human life would terminate in a rational, liberal-democratic, and capitalist order standing against a scattered and dispersing set of enemies.

Fukuyama's argument was based not just on philosophical speculation, but on a reading of then-current events: the collapse of communism, the flourishing of electronic media, the apparently frictionless opening of borders, and a stock market beginning an epic bull run.

Today his thesis seems like a monument to the dreams of an earlier era (one chapter was titled "The Victory of the VCR"). Our cultures are evolving today, but not, it seems, toward any harmony. The chaos of the 21st century makes our simulations feel immediately familiar. Two decades after 9/11, even the Western liberal democracies are willing to consider dark models of human behavior, and darker theorists than Fukuyama.

**A NEW YORK CITY SHIBBOLETH:** Local New Yorkers can quickly identify a tourist through his or her mispronunciation of this downtown street name.

lillisphotography / istock

Carl Schmitt, for example, who saw the deliberative elements of democracy as window dressing on more authoritarian forms of power. Or Robert Michels, whose studies of political inequality led him to see democracy as a temporary stage in the evolution of society to rule by a small, closed elite. As intellectuals at both political extremes increasingly see the possibility of a rational political order as a fantasy, Shibboleths take up their role in defining racial, national, and religious boundaries and appear once again to be ineradicable features of political life.

There is a great, and rich, valley between these philosophies, and another between the computer models that match them—between the simple, violent and less-than-rational agents that John Miller and I simulated, and the super-rational cooperators that Fukuyama might have considered to be waiting at the end of history. The models, at least, encourage a guarded optimism.

Researchers associated with meetings at the Machine Intelligence Research Institute (MIRI) in Berkeley have studied the behavior of rational but resource-limited machines who could inspect each other's source code. Such transparency might seem to solve the problem of cooperation: If I can predict what my opponent will do by simulating his source code, I might decide cheating is not worth the cost. But what if my opponent's code includes a simulation of what I will do as a consequence of running that simulation, and tries to exploit that knowledge? Without the symmetry of perfect rationality, this problem leads to some extreme mental contortions.

Some of the machines in MIRI's bestiary might remind you of people you know. “CliqueBot,” for example, simply cooperates with anyone who shares the same source code. It only cares about codes that match its own letter-for-letter. “FairBot,” on the other hand, tries to look beneath surface differences to prove that an opponent will cooperate with someone like itself. Informally, FairBot says, “if I can prove that my opponent will cooperate with me, I'll cooperate with him.”

How do these machines get along? While the full solution is a paradox of regress, studies of predictive machine behavior in a Prisoner's Dilemma standoff provide the comforting answer that mutual cooperation remains at least possible, even for the resource-limited player. FairBot, for example, can recognize similarly-fair machines even if they have different source code, suggesting that diversity and cooperation are not impossible, at least when intelligence is sufficiently high.[1]

Even the genocidal machines at the violent end of the spectrum may carry a heartening lesson. They emerged from the depths of a circuit board, simulated on a supercomputer in Texas. They had no biological excuse to fall back on. Maybe we, too, shouldn't make excuses: If a behavior is so common as to emerge in the simplest simulations, perhaps we ought neither to fear it, nor to idolize it, but to treat it, the same way we do cancer, or the flu.

What if we saw tribalism as a natural malfunction of any cognitive system, silicon or carbon? As neither a universal truth or unavoidable sin, but something to be overcome?

*Simon DeDeo is an assistant professor at Carnegie Mellon University, where he runs the Laboratory for Social Minds, and external faculty at the Santa Fe Institute.*

*The author would like to thank the Alan Turing Institute for their summer hospitality while this article was written.*

### References

1. Barasz, M., *et al*. Robust cooperation in the Prisoner's Dilemma: Program equilibrium via provability logic. *arXiv* 1401.5577 (2014).

JOIN THE DISCUSSION

**17 Comments** **Nautilus** 1 **Login**

♡ **Recommend** 4 **Share** **Sort by Best**

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

**Steve SanFrancisco** • 7 days ago

Tampering with nature and overcoming the tribalism that emerges as a natural property of a self organizing system is questionable. How would you tamper with something that fullfills a purpose the author doesn't seem to understand? Why would you tamper with something you don't undertstand? I suggest you run the simulation with your tampering added and see what happens.

0 ^ | v • Reply • Share ›