The Ethical and Social Ramifications of Rossler's Blueprint for Making Benevolent Robots

Kevin LaGrandeur, Ph.D.

Faculty, NYIT; and Fellow, IEET

Digital systems made to enhance human physical capabilities as widely diverse as caretaking and killing are becoming increasingly prevalent. These digital helpers are, in turn, rapidly becoming more complex and independent: some computer specialists think that we are on the verge of having what computer scientists refer to as "strong AI"—Artificial Intelligence that can learn to evolve and to think on its own. Good examples of networked devices that hover on the edge of this potential are the "caretaking" robots being developed by the Japanese to care for the elderly and "smart" weapons being developed by the military, such as drones.

But the whole scenario described above has scientists and some philosophical commentators worried. What is to prevent strong AI from becoming dangerous to humans, to prevent a condition like that described most radically in the *Terminator* film series, where the defense AI called "Skynet" decides to wage war on its human creators? (There have already been some documented problems that eerily foreshadow this kind of possibility, such as the infamous malfunction of a smart antiaircraft gun in South Africa in 2007, where it turned and began killing the soldiers operating it.)

One answer is that posited by the German Complexity Theorist Otto Rossler. He maintains that AI can be programmed for benevolence. Applied ethical theories of this kind originate with Leibniz and Hume, and Rossler uses these as a precedent. He mainly, though, focuses on more modern theories of Spatial Darwinism and of social bonding of the type made famous by Konrad Lorenz. Combining these with mathematics, Rossler claims that benevolence is possible to program, since the brain can be effectively modeled using differential equations.

But would this kind of theory work? I see some key problems with the bridge he proposes between bonding theory and spatial Darwinism on the one hand, and benevolence on the other. If his model holds water, would it even create a "benevolent" AI, or just an AI with a varying sense of "attachment" to a particular operator? And would this, in turn, just risk greater nuisance springing from the unintended consequences of such attachment? My presentation will address these and other ethical issues related to his theory.