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

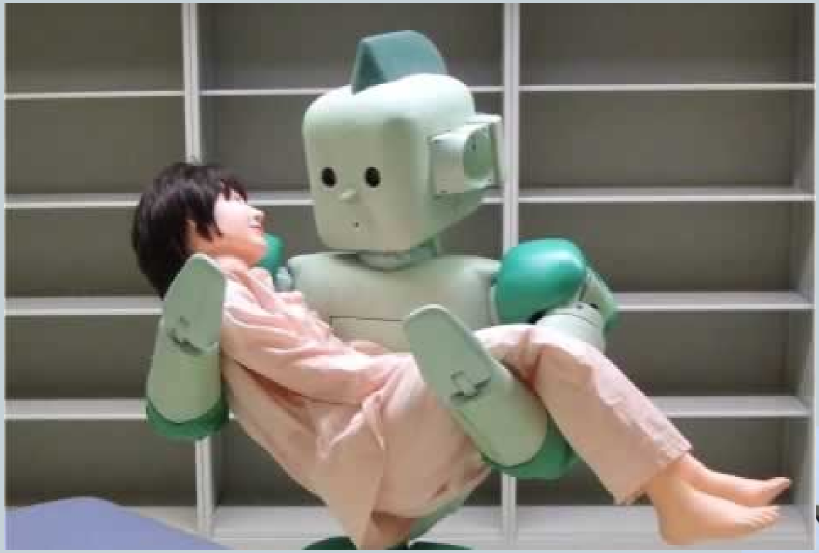
KEVIN LAGRANDEUR, PH.D.
NYIT

Our dreams of artificial servants are not new…

- My book, *Androids and Intelligent Networks in Early Modern Literature and Culture: Artificial Slaves* (Routledge, January, 2013), shows
- They stretch back to *The Iliad*, Aristotle's *Politics*

Today, increasing dependency on artificial proxies for things as diverse as

- Caretaking



www.businesspundit.com

- And Killing

Is ever-stronger AI on which we increasingly depend safe?

- The *Terminator* series: an extreme scenario, but not unreasonable

- Computer Scientists worried enough to debate limits on AI research in 2009.

An analogous, real-life catastrophe

- The South African autonomous anti-aircraft cannon that killed 9 friendly soldiers (Wired Magazine, Oct. 18, 2007)

One answer: build benevolence into AI

- Benevolence and AI: Otto Rossler (German physicist and complexity theorist). References:
  - “Nonlinear Dynamics, Artificial Cognition and Galactic Export.” (2004)
  - “Delictatio in felicitate alterius—Benevolence Theory.” (2004)
  - *Neosentience : The Benevolence Engine* (co-authored with Bill Seaman). U of Chicago Press, 2011.

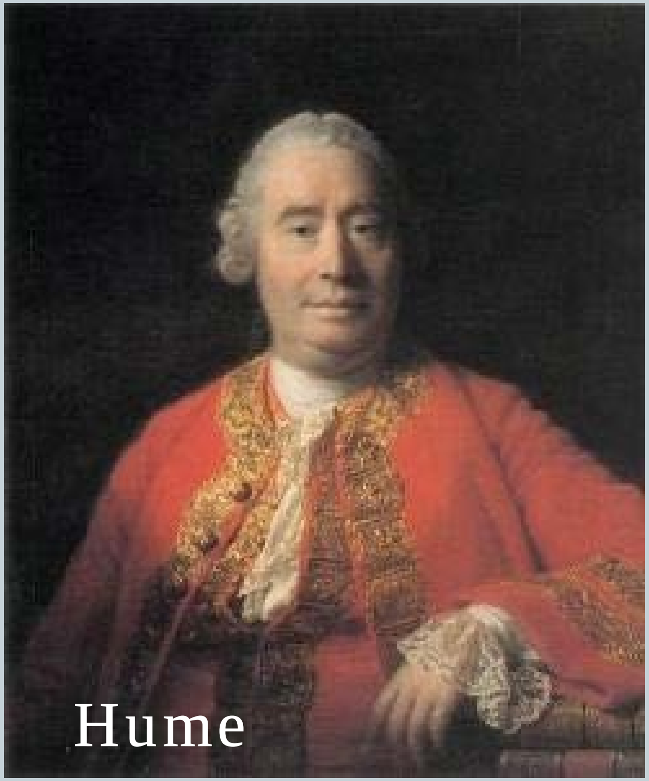
Benevolence theory

- “morally valuable character trait—or virtue—of being disposed to act for the benefit of others” (Stanford Encyclopedia of Philosophy)

Benevolence theory

- Origins:

Leibniz



Hume

Leibniz’s Basic Idea

- Leibniz: “God's universal benevolence...is an ideal we ought to do our best to imitate and continuously aspire to. The more one's benevolence expands to encompass the happiness of...others, the more one grows in justice and virtue, thereby increasing the moral good.”
- Rossler’s idea for translating this into programming: combine “Spatial Darwinism” and...

Social Bonding Theory

- (Made famous by Konrad Lorenz)

Formal mathematics can be found here:

- Rossler, O. E., “adequate locomotion strategies for an abstract organism in an abstract environment: A relational approach to brain function”; *Lecture Notes in Biomathematics*, vol. 4, 342–369 (1974).

Specifics: Social Bonding

- Adaptive survival trait, so
- “Programmed” into the neural makeup of animals
- Since AI functionality is based on animal brains, and
- Since brain functions are reducible to equations (I’ll explain in a minute),
- Bonding could be programmed into a machine.

How? Via Spatial Darwinism

- First, the mathematical basics:
- A “brain” = combination of two differential equations:
  - First = an “autonomous direction optimizer”

Second Equation

- Describes “virtual reality generator”…

for building an internal picture of the outside world

First + Second Equations = “Autonomous Path Optimizer”

- Determines where to go and when, and in which order, so as to optimize the **individual** organism’s survivability in the **short run**
- **= Rossler’s** “spatial Darwinism”
- (versus Darwinism, which deals with the survivability of the **species** in the **long run**)
- Direction optimization reaction to positive or negative potentialities, so parallel to emotions

The Way this Works in Humans

- Mother’s smile bonds the infant

Mother shares a piece of apple with infant

Which gives the infant an idea…

- Infant tries to extend experience of bonding-smile by offering mother a return piece of apple

Mother smiles at toddler’s offer

- baby experiences the unexpected: Leibniz’s pleasure in the joy of the other

Results:

- Leads to further benevolence by infant, and
- A projection of an “Other over there” in the mother
- Chain reaction continues

Machine Analogue to this Benevolent Experience

- Rossler's mathematical models define "optimal path" as staying near one human
- who = Lorenz's "animal with home-valence"
- Which = mother figure, in most animals

Rossler’s Posited Results

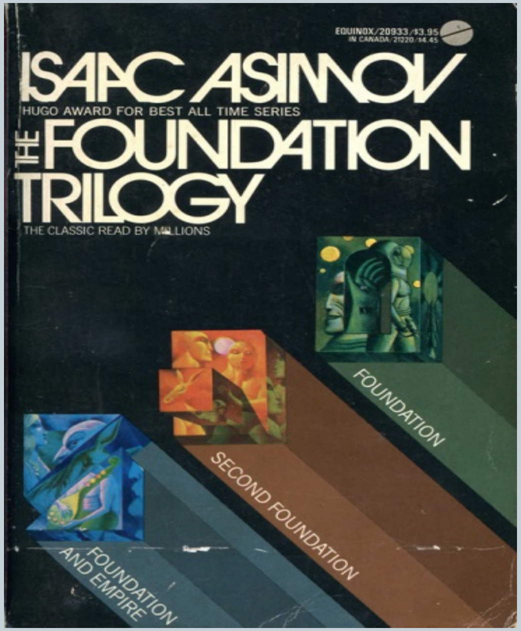
- This is an attachment algorithm based on proximity
- The human would (I presume) develop attachment based on the machine’s “loyalty” (it follows her and stays close).
- Rossler: The machine would “learn,” just like the baby, that offering things to its human would provide positive feedback, making a feedback loop that would self-perpetuate

Ethical/social considerations

- Conceptually, this model does not differentiate between bonding and mere proximity
- The emotionally-laden valence-bonding may not be so easy to produce (or learn, for the AI)
- Would it create true benevolence, in the philosophical/social sense?
- Ethics of creating suffering in a super-intelligent AI?

Unintended consequences of success

- Would AI benevolence definitions = human definitions?
- AI evolution may = evolution of goals and norms “alien” to ours



Conclusion

IN SUM, I SEE THE IDEA OF BENEVOLENCE IN MACHINES AS A NOBLE GOAL, I’M JUST SKEPTICAL ABOUT ITS PRACTICALITY, GIVEN THE LIMITS OF OUR PROGRAMMING CAPABILITIES, THE CURRENT PROSPECTS FOR SENTIENT AI, **AND THE FACT THAT THE PROGRAMMING CONCEPTS THAT ROSSLER SUGGESTS LEAD AT BEST TO AN** ***IMITATION*** **OF BENEVOLENCE.**