Thirteen Theses on the "Control Problem" in Superintelligence:

A Commentary by William Leiss
www.leiss.ca

30 June 2016

Note to the Reader:

I intend to develop these preliminary comments into a larger essay, and I would appreciate having feedback from readers of this document. Please send your comments to:
wleiss@uottawa.ca

Contents:

A Short Introductory Note on the Idea of Superintelligence

The idea of a "superintelligence" (or superintelligent entity) is best-known from the book of that title by N. Bostrom (University of Oxford), published in 2014. It is defined by Bostrom as an entity that has mental powers far beyond those of any human being in terms of general intellectual skills, especially in cognition, memory, and recall, and possibly even in social skills. This idea was originally conceived as long ago as 1956, with the contention that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

This idea is broader than just machine-based simulation, and it is now suggested that there are a number of different ways in which such an entity might be brought into being by human action, including:

1. Artificial (machine) computer-based intelligence (AI) operating at a very high level of performance, based on algorithms utilizing neural nets and "deep learning," the circuitry of which runs at close to the speed of light, and thus orders of magnitude faster than the biological circuitry of the human brain.

2. "Whole-brain emulation," that is, scanning and thus digitizing an entire human brain, encompassing the entirety of its 100 billion neurons and all of their synaptic connections, and then manipulating the result to increase this artificial brain's cognitive and other powers;

3. Selective breeding of humans who have been genetically-engineered to have superior brain functions. Obviously this could only be carried out across a number of generations and lifetimes.

Networking many individual entities of the first two types above could also produce levels of performance that are orders of magnitude higher than any single entity operating alone.

These discussions immediately raise the key issue, which is known as the "control problem": If any entity exhibiting superintelligence were to come into being through human design, would it be strictly subordinate to human wishes and commands, for all time to come, or could it in effect "escape" and develop an autonomous will and mode of action? Would it thus become "self-conscious," that is, aware of itself and of its powers?

In other words, could it "escape" from human oversight and control by deceiving its human minders about what it was up to – pursuing what Bostrom calls (p. 155) "clandestine goals" – until it was powerful enough to resist their attempts to bring it back under human control? If so, would it then "wish" to dominate us or perhaps even exterminate us? (See Chapter 7 of Bostrom's book, "The Superintelligent Will.")

Clearly, if we were to seek to design and build such an entity, we would be doing so in search for benefits for humanity which far exceed what we can now derive from our ingenuity, work, and technologies. As in any such development, ideally some public authorities would want to do a risk assessment on such an entity – presumably well in advance of "switching it on" – and quantifying its associated upside and downside. Bostrom's book is excellent in describing what I call the "downside risk" in such situations, that is, the bad things that may result and may make us regret our wishing to create such an entity.

And there are some very bad possibilities indeed, which Bostrom refers to as "existential catastrophes" for humanity (pp. 218ff.). These are what I have referred to elsewhere (in my 2010 book, *The Doom Loop in the Financial Sector*) as "black-hole risks." A black-hole risk in one where we cannot even calculate properly, in advance, either qualitatively or quantitatively, the dimensions of the downside risk, in terms of lives lost and economic collapse, and thus there is the very real possibility that we would have little or no idea what might happen, and no chance to reverse the course of action once it started to unfold.

But do we even have – collectively, as humanity – the power and authority to decide, in some equitable fashion, whether we want to proceed towards creating such an entity? If global technological progress approaches the point where it seems increasingly feasible to make the attempt, and decide against it, do we have the capacity to ensure that some rogue nation or super-wealthy private entrepreneur cannot be stopped from proceeding?

Bostrom's book is brutally frank and honest in suggesting that *we may ultimately fail to solve the control problem.* Thus, I suggest, it's time to begin discussing this issue in broad public forums. I have presented some of my initial thoughts in the following pages, and I invite you to respond with your own.

Postscript: The discussion about MI (machine intelligence) – a descriptor I prefer to that of AI ("artificial intelligence") – was stimulated recently by the victory of AlphaGo, a computer game-playing program developed by Google's DeepMind project, over the reigning human world champion, South Korea's Lee Sedol, in the ancient game of Go. AlphaGo defeated Sedol by 4-1 in a five-game match, a victory many experts thought would take much longer to occur. In Game Two, at Move 37, AlphaGo made a winning move that stunned Sedol and the other championship players who were watching, one of whom described it as "so beautiful." But then, in Game Four at Move 78, Sedol flummoxed AlphaGo with a winning move of comparable creativity. The expert commentaries strongly suggested that each of the contestants had learned something new about the game of Go as a result of playing each other.

William Leiss
30 June 2016

The Control / Self-Control Problem in Superintelligence: Thirteen Theses

1. The "control problem" is, at a deeper level, the "self-control problem":
    - "Control" implies externally-imposed, "self-control" internally-generated.

2. The definition of intelligence is "instrumental rationality (IR)":
    - *Superintelligence,* p. 217, "convergence on instrumental values."

3. Relevant here is Max Weber's typology of rational action:
    - The two most importance types are *Zweckrationalität* ("instrumental" or "means/ends" rationality and *Wertrationalität* ("value" rationality).

4. There appears to be a fatal flaw built into the conception of the control problem: Value-rationality is to be imposed atop instrumental rationality – and this is unlikely to work:
    - The discussion in *Superintelligence* strongly implies that "the control problem" is a "hard" problem – like the unsolved problem of consciousness – and may in fact be insoluble as presently posed (unable to avoid deception, etc.).

5. The order of priority as between the two forms of rationality should be reversed: Value Rationality – operationalized in algorithms – must be the foundation-stone of any superintelligence. In other words, the superintelligence agency itself must be, first and foremost, a *moral agent*, and must be such before it reaches the point of autonomous breakthrough.
    - In other words, a control system – more precisely, a self-control system – must be internalized in the structure of its operating routines.

6. Any and every being – and more strongly so for any superintelligent being – which can regulate its behavior autonomously, by independently *willing* a course of future action, and which is aware of this capacity (consciousness?), may be regarded as a "self" or an "I."

7. The First Principle for any Superintelligent Being should be: It is unwise, and almost certainly catastrophic, to create any agent possessing superintelligence without *first* being sure (to a very high degree of probability) that it would unfailingly exercise *rational self-control* in some meaningful sense:
    - In other words, its instrumental values must be subordinated, in its decision routines, to *its own* ethical value-system that is demonstrably governed by human values (as defined and operationalized, see below).

8. Failing this assurance, it is rational for humans to oppose strenuously the creation of any agent with superintelligence:
    - At least one capable world superpower (either the UN or some state) must announce that it is willing to use weapons of mass destruction to prevent any

superintelligence that is not an autonomous moral agent, governed by humane values (as above), from coming into being.

9. Further, any superintelligence ought to be construction so as to be *better* (in moral or ethical terms) than the deeply-flawed humanity it is meant to serve (Kant's "crooked timber"); otherwise, what would be the point? (See my short story, "Good Robot.")
    - Just look around us at the state of the world: Why would anyone in his or her right mind want to create a superintelligent agent that would not be reliably and demonstrably *superior* in its decision routines, *in an ethical sense*, to the mass of humanity (and its leaders) at the present time?

10. An ethical foundation for the operation of self-control in any superintelligence should be built on the basis of humanity's best effort at creating an overriding set of ethical norms, combining:
    - The Hippocratic principle: "First do no harm."
    - The Kantian Categorical Imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction."
    - The Golden Rule: "Do to others what you want them to do to you."
    - Justice as Fairness (Rawls) and its sub-principles.
    - "Serve the people."
    - Others to be added as necessary and appropriate.

11. The idea of self-control also needs to be further characterized and then operationalized in an *evolutionary* context, as impulse control, self-regulation (the Freudian superego), delayed gratification, empathy, etc. (already evident to some extent in chimpanzee behavior, arising out of intense sociability), and then combined with the list of ethical norms, above:
    - Do a thought-experiment: Imagine what human society would be like if there were no "natural" self-control elements, ultimately built into our genome and then into our developing brain!

12. Self-control in at least most human agents arises innately, "naturally," or spontaneously as a result of their evolutionary heritage; severe deficiencies in innate self-control in such agents (correlated with deficits in specific regions of the brain) are reasonably regarded as being *pathological* and associated with some forms of serious criminality.
    - It follows that any self – human or machine – having no *innate* mechanism of self-control should be regarded as being pathological.

13. In sum: If we cannot solve the self-control problem, we will *never* solve the control problem with sufficient reliability to justify creating an autonomous superintelligent agent.

Future Social Scenarios and the Control / Self-Control Problem [C/SC]

1. My contention is that the C/SC problem needs to be addressed for humanity *before* the creation or near-creation of superintelligence.

2. The C/SC theme in political theory begins at the beginning, with Plato's *Republic* and its conception of justice.
    - To some extent Plato's formulation is countered by the famous question from the Roman poet Juvenal: "Qui custodiet ipsos custodes?"

3. Background to the control problem in the "consequences of technological progress" theme in nineteenth-century literature and in the technocracy movement of the 1930s:
    - See my essay, "Sublime Machine" (1985; *Under Technology's Thumb,* 1990).

4. And in the dystopias, especially the two earliest ones, E.M. Forster's great short story, "The Machine Stops" (1909), and Yevgeny Zamiatin's amazing short novel, *We* (written 1919).

5. My futuristic scenario deals explicitly with this problem:
    - Vol. 1, *Hera, or Empathy* (2006); vol. 2, *The Priesthood of Science* (2008); vol. 3, in progress; this project was started in 2002.
    - Thematically, it begins with Francis Bacon's problem: How will the great powers about to be conferred on humankind by modern science and technology be superintended? His answer was: by religion and its value-structure.
    - The hope of the French Enlightenment (FE), on the contrary, was that the scientific method itself would play this superintendence role, by gradually diffusing through society and displacing religion, superstition, and savagery.
    - My trilogy deals with the failure of both Bacon's scheme and that of the FE; in it, an elite group decides to "hide" science from ordinary humanity – on account of the willingness of humans to use every means of destruction in their ongoing murderous ways – and by so doing seeks to perpetuate the scientific ethos and to ensure that its outputs are used only for humane ends.

6. Where volume 3 of my trilogy is headed:
    - The last, long section is entitled "Dialogues concerning the Two Chief Life-Systems" (reference to a very famous book at the beginnings of the scientific revolution).
    - The two life-systems are silicon and carbon; there is an extended discussion – set 50 years or so in the future – between the leader of my elite human group and a superintelligent machine agent which they have allowed to become autonomous.
    - In other words, this machine agent, with its internalized self-control feature, has been allowed to come into being *after* human agents have solved the general self-control problem of humanity itself.

