Three Observations on Artificial Intelligence

1. *We are They*

   Francis Crick called it the "Astonishing Hypothesis": that consciousness, also known as Mind, is an emergent property of matter. As molecular neuroscience progresses, encountering no boundaries, and computers reproduce more and more of the behaviors we call intelligence in humans, that Hypothesis looks inescapable. If it is true, then all intelligence is machine intelligence. What distinguishes natural from artificial intelligence is not what it *is*, but only how it is made.

   Of course, that little word "only" is doing some heavy lifting here. Brains use a highly parallel architecture, and mobilize many noisy analog units (i.e., neurons) firing simultaneously, while most computers use von Neumann architecture, with serial operation of much faster digital units. These distinctions are blurring, however, from both ends. Neural net architectures are built in silicon, and brains interact ever more seamlessly with external digital organs. Already I feel that my laptop is an extension of my self – in particular, it is a repository for both visual and narrative memory, a sensory portal into the outside world, and a big part of my mathematical digestive system.

2. *They are Us*

   Artificial intelligence is not the product of an alien invasion. It is an artifact of a particular human culture, and reflects the values of that culture.

3. *Reason Is the Slave of the Passions*

   David Hume's striking statement

   > Reason Is, and Ought only to Be, the Slave of the Passions

   was written in 1738, long before anything like modern AI was on the horizon. It was, of course, meant to apply to human reason and human passions. (Hume used the word "passions" very broadly, roughly to mean "non-rational motivations".) But Hume's logical/philosophical point remains valid for AI. Simply put: Incentives, not abstract logic, drive behavior.

   That is why the AI I find most alarming is its embodiment in autonomous military entities – artificial soldiers, drones of all sorts, and "systems". The values we may want to instill in such entities are alertness to threats and skill in combatting them. But those positive values, gone even slightly awry, slide into paranoia and aggression. Without careful restraint and tact, researchers could wake up to discover they've enabled the creation of armies of powerful, clever, vicious paranoiacs.

   Unlike in the case of nuclear weapons, here there are no clear and obvious red lines.

Incentives driving powerful AI might go wrong in many ways, but that route seems to me the most plausible, not least because militaries wield vast resources, invest heavily in AI research, and feel compelled to compete with one another. (In other words, they anticipate possible threats and prepare to combat them ...)

How might we avoid that danger, while reaping the many rewards that AI promises? I think transparency and open discussion is essential. The Wikipedia and open source programming communities provide inspiring examples of openness, in closely related endeavors. Their success demonstrates that very complex development projects can thrive in an open environment, where many people keep careful watch on what's happening and maintain common standards. It would be an important step forward if AI researchers pledged, collectively, to abstain from secret research.