IMusing · 001

Other Cognitive Agents — Skeleton

Brian Cantwell Smith
January 14, 2017

I. Problematic

- *A.* AI on the verge ("Some day soon") not just of specific intelligence (reading x-rays, playing Go, driving cars, translating text, managing power electrical distribution systems, and so on), but of *general cognition*
- B. Humans no longer be the only general rational systems.
    1. Animals are arguably rational in some respects—exquisitely tuned, in many cases. But they are not general rational agents. Not just lack of language, but perhaps related,
- C. Raises the question of how we should live with them—appropriately, synergistically, ethically.
- D. To answer that requires *understanding what these systems will be, do, and be capable of*.
    1. How to live with them
    2. How to build them
    3. How to anticipate, avoid, and interpret challenges and problems that arise, and
    4. What to hold them accountable to
- E. Not an isolated development, in some Frankenstein lab. For imaginative purposes, the most compelling image may be of a single "smart" computer, like Kubrick's legendary Hal. Realistically, the prospects of artificial cognition rest on radically networking, so that it is more that there will be networked capacities of rationality, cognition, etc. Even: Siri, Alexa, driverless cars.

II. "Split Realm" View

- A. Tempting way to view this future—roughly continuous with how we implicitly use and understand computational technology to date (and not so Δ from how we view all technology)
    1. Outsource (a lot of our) *rationality* to machines of our devising;
    2. Reserve to humans "deeper" properties—ethical judgment, emotions, etc.
    3. I.e., Copernicus (universe), Darwin (creation), us (now) dethrone us from our cherished position as centre or pinnacle of *intelligence*.
    4. Is it over for us? Have we created our natural successors (even: replacement)?
    5. No, the argument would go. Rationality, at least rationality alone, is not what matters.
    6. Rather, what remains uniquely human: ethics/values/judgments (higher emotions)
- B. Remarks
        - a. Start with the "sanguine" position: outsource "rational thought" ('calculation,' we might call it—along with display, control, etc.). Continuation of where we are. Yes, might be discontinuous in *power*, but not in conception. And perhaps less total in its impact that the triumphalist would expect (this is Meg & Jill's position).

2. Very comfortable (not the future, but the view)
3. Cite Friedman
    - i. Friedman's "From Hands to Heads to Hearts"[1] (NYTimes, jan 4, 2017)
    - ii. "Therefore, Seidman added, our highest self-conception needs to be redefined from "I think, therefore I am" to "I care, therefore I am; I hope, therefore I am; I imagine, therefore I am. I am ethical, therefore I am. I have a purpose, therefore I am. I pause and reflect, therefore I am.""
    - iii. "Machines can be programmed to do the next thing right. But only humans can do the next right thing."
4. Answers lots of questi0ns
    - a. Maintains a clear Δ
    - b. Because we are arbiters of what matters, aren't metaphysically challenged
    - c. *We* matter—our concerns matter most, humans must be valued, etc.—because we are the locus of *worth*.
        - i. fn: Humans matter "because." ⟸ no good. There must be something about us *in virtue of which we matter*—something that makes it the case that generating life is one of the best things we do, and taking it the worst.
    - d. To what do we hold them accountable? Our goals,
    - e. Do *they* matter? Only instrumentally. (Don't have to worry about unplugging them)

C. Call it the "split-realm" view

1. Can sequester the rational from the ethical

D. Failure

1. Nothing in the two-realm view ensures a happy future.
2. Could be disasters, of various sorts
    - i. Human mismanagement: Military uses, surveillance, destruction of privacy, neocapitalism gone wild, etc. Also: excessive technological enthusiasm, at the expense of human values. Bad, yes, but again: familiar, from prior technologies.
    - ii. Geoff's position, basically: that the systems will self-organize, respond to natural competitive pressures. Need a scenario
        - a. Maybe driving. Driverless cars; accident minimization; interfere with licensing; deny licenses to people with bad records; then old people; then people at all. Allocation of transport to more "worthy" customers. Ultimately, system control of all human movement. Then: control of delivery, etc. Essentially a 1984 vision, but ruled by computer systems, rather than a political elite.
        - b. Another: VLSI retinas. At first, for the blind. Then: reduce insurance. Then: prohibit driving at night by people with normal retinas. (Maybe this could figure into the above.)
        - c. Perhaps a third: military systems.
    - b. Human

[1] http://www.nytimes.com/2017/01/04/opinion/from-hands-to-heads-to-hearts.html?mabReward=R7&recp=2&moduleDetail=recommendations-2&action=click&contentCollection=Politics&region=Footer&module=WhatsNext&version=WhatsNext&contentID=WhatsNext&src=recg&pgtype=article

- i. We subserve them to greed, or capitalism, or war, and jealousy, or racism, or xenophobia
- ii. Or innocent of malicious intent, we exaggerate our sense of what they might be capable of (e.g., replace health care workers, and evacuate medical treatment of "care").
- iii. Joy
  - a. (Kacysnki): "we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions."
  - b. (Joy) "robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate
  - c. (JSB & Duguid) "All forms of artificial life (whether bugs or bots) will remain primarily a metaphor for—rather than a threat to—society, at least until they manage to enter a debate, sing in a choir, take a class, survive a committee meeting, join a union, pass a law, engineer a cartel or summon a constitutional convention."
  - d. (JSB & Duguid) "Malthus and Wells helped prevent the very future they were so certain would come about."
  - e. (JSB & Duguid) "Social and technological systems do not develop independently; the two evolve together in complex feedback loops, wherein each drives, restrains and accelerates change in the other. Malthus and Wells—and now Joy—are, indeed, critical parts of these complex loops."
  - f. pass a law, engineer a cartel or summon a constitutional convention.
  - g. MJ's is negative as well—but not doomsday in the same sense (that the machines will take over). Geoff's argument has the air of inevitability. M&J attribute the "ill" to people—their fetishization of profit, efficiency—and failure to acknowledge impacts on jobs, etc.
    - — That is: M&J basically believe, I think, that the technology *could* be built in such a way that all would be well, but that our failure to do so simply reflects human failure in its construction . stewardship.
    - — One way to put this might be this: that M&J think that the sanguine option is in some sense eminently feasible; that anything else is human failure. I just asked Jill, who confirmed what I thought: that her sense is that this would not so much *curtailing* or *limiting* what is developed, but deploying it, stewarding, devoting it, etc., to good purposes.

c. Machinic

- i. They take over, for any number of reasons
  - a. Darwinian survival; uncontrolled obedience to programs, without any care
  - b. there will be a competitive/Darwinian struggle for resources, and that we will lose—all without attributing values, other than greed/selfishness and struggle for survival, to the machines.
- ii. Example: drivers, …

d. Disruption

i. Unleash job displacement, trigger revolts, etc.

ii. Something else: the labour disruptions, etc. (Stefan's view, maybe—or anyway I can label it that way here) is that there will be disruptions, maybe for 100 years, but that always happens, and that we will adapt, and everyone will be better off afterwards.

a. I'm wondering if this could be labeled a positive (+) reaction to the disruption point, whereas one could have a negative (–) reaction to it: there will be rebellions, and maybe fundamentalist religious take-overs, etc.

b. Note that Geoff didn't consider the disruption point at all.

c. Do I want to cast this as a "cartography of responses"? Or write it as a "fortunately, unfortunately" story. Tai!

E. Still: the split-realm view could be comfortably deployed in *analysing* these eventualities.

1. Basically

F. Unfortunately, I think it is wrong.

III.… consider …

a. These scenarios, both good and bad, involve systems whose *behaviour falls under ethical norms* (i.e., they do good or bad things)—but roughly in the way in which any technology would be so governed (chemical weapons, the electrical power grid, etc.). That is, they are considerations of ethical uses.

They are not imagined—at least nothing in the discussion so far has considered them—as *themselves ethical agents*, in the sense of engaging in ethical deliberations. They may have norms built in, that is—e.g., avoid accidents, maybe even Asimov's laws of robotics[2]—but are not (yet) being imagined as ethical agents —not as *reasoning about ethics per se*. [not a black-and-white Δ]

That is: the discussion so far just deals with computational things as *technology*.

b. Nor do they involve any consideration of the systems *themselves as ethical* (e.g., of whether they are themselves ethically important (other than being instrumentally important, in the way that the power grid is instrumentally important). That is, they are not yet considered to be ethical subjects.

c. Moreover, nothing has been said, yet, that is specific to the original hypothesis: that these systems will be capable of (something like) *rational thought*—i.e. we could say, that we are imagining the day when computational devises of our own consideration will be rational agents.

Actually, this may not be true: it is possible that their capacities for "thinking" may figure in Geoff's response».

IV.Rationality

A. Problem: it doesn't address what rationality actually *is*.

[2] (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
(2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
(3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws

   a. « Make a point about the achievement of rationality is huge. One of the most important facts is that a rational creature can *decide* [fn. on 'decision']. But not just decide; involve wide ranging considerations, with far-reaching impact. That *range* is critical.

   b. Also, one reason that it can have these wide-ranging considerations is that they involve the use of *representations*—which represent distal facts with local (proximal, effective) structures. And (the essence of rationality) they behave in normative accord with those distal states of affairs. Of course books, famously, represent as well —but books can't *do* things appropriate to what they represent. Data bases do too—and system using data bases do act in accord with the "c0ntent" of their representations. But not *arbitrarily*; they do things pretty much as specified by their programmers and designers. *Rational* systems, on the other hand, will not be so constrained—their generality is a mark of their "intelligence" [fn.: cf. Descartes' talk of generality]

2. Underestimates the magnitude of the achievement. and the gravity of the consequences

   a. Maybe say: it will be stunning, when it happens, for us to have conversants who genuinely engaged in discussions …

   b. Won't serve as an adequate "imaginary" to allow us to design systems appropriately, predict or comprehend the consequences we may be unleashing, or supply sufficient analytic power to deal with situations that arise.

      i. No serious treatment of the assumption that these systems will be able to "think for themselves"—or indeed to deal with what it is for these systems to *think*, as opposed merely to acting (behaving) intelligently in various task domains.

      ii. No consideration of the systems *themselves as ethical* (e.g., of whether they are themselves ethically important (other than being instrumentally important, in the way that the power grid is instrumentally important). That is, they are not yet considered to be ethical subjects.

      iii. Moreover, nothing has been said, yet, that is specific to the original hypothesis: that hese systems will be capable of (something like) *autonomous rational thought*—i.e. as rational agents.

3. But mostly I want to argue that the idea that rationality can be sequestered from ethics isn't true. That means that, if we build systems capable of genuine thought, they will effectively embody an ethical stance, whether we like it or not. And not just an implicit ethical framework; they will necessarily end up engaging in explicitly ethical deliberation. So the questi0n is going to be how we provide them with a worthwhile ethical framework.

   a. That is: split-realm view assumes we can *sequester the ethical away from the rational* (and the *rational away from the ethical*).

   b. One lesson from the philosophers who have thought hardest about these issues is that that is not possible.

4. No dealing with them as *ethical agents*, in the sense of *engaging in ethical deliberations*. They may have norms built in, that is—e.g., avoid accidents, maybe even

Asimov's laws of robotics[3]—but not (yet) imagined as *reasoning about ethics per se*. [not a black-and-white Δ]. That is: the discussion so far just deals with computational things as *technology*.

a. Why will they need to deal with ethics as an explicit subject matter (i.e., use predicates like 'good' and 'better' and 'bad' in their rational deliberations—not simply "how far away" or "how much power".
b. Think of Asimov's first law: "A robot may not injure a human being or, through inaction, allow a human being to come to harm." Untenably simplistic. Administering a drug with side effects (that is, just about any drug) violates the "do no harm."
c. Not just *think about doing the right thing* (e.g., avoiding accidents, or poisoning people)
d. They will have to *think about what is the right thing to do*.

B. Features of rationality

i. General considerations—and general subject matter (wide-ranging)
ii. Committed to truth—to getting things right
iii. Involves the use of representation (because you can't always check with the world)
iv. Relation to the world (reference) is not causal
    a. Representational mandate
v. Result: discontinuous from what we are used to (≠animals)

2. There are five features of rationality that will matter most here.
3. Thinking has three distinctive properties
    a. It is *world-involving*
    b. It traffics in *meaning*
    c. Semantics, at least, and perhaps meaning as well, is *not a causal phenomenon*.
    d. It is committed to *truth*—to "getting the world right."
        i. This is where deference comes in?
4. To get at it, need to know more about cognition and rationality
    a. Committed to *getting things right*. E.g., intelligent utility controller: needs to be *correct* about things: power plants capacity, impact of weather, etc.
    b. That is: can't be rational without the world. Need to be able recruit the world to their projects.
    c. That is: need to be committed to the *truth*. Doesn't mean they can't be horribly instrumentalist. But knowing what will work, instrumentally, requires knowing what is the case. If you are going to turn off a power plant, because of a lull in the energy demand, then you have to be right that there *is* a lull in the energy demand.
5. That is: deference, non-causality, representations (writ large).
6. Read Joy…

C. Now it is against this background that we need to assess the presupposition behind the

[3] (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
(2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
(3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

split-realm view

V.Ethics

1. No consideration of the systems *themselves as ethical* (e.g., of whether they are themselves ethically important (other than being instrumentally important, in the way that the power grid is instrumentally important). That is, they are not yet considered to be ethical subjects.
2. My point, I suppose, is that they need to be given an ethical framework, to which they are held accountable—*by themselves, and by each other*, not just by us. Or anyway maybe that is my point.
   a. And perhaps I want to argue (not sure!) that if one gives them a normative framework strong enough to ensure that they hew to the truth, they will also hew to the good? Boy, that would be nice. And *ultimately*, I think I believe it. But in the nonce? I doubt it…
3. Re ethics
   a. … this stuff is still weak … ;-(
   b. Ethical considerations (a long way from the "take it out and boil it"), but the point remains:
   c. Themselves committed (truth is normative)
4. …
5. These facts undermine any thought that they aren't:
   a. Ethical *agents*
   b. Ethical *subjects* (worthy in their own regard)
6. Bodies
   a. What about *embodiment*?
   b. What about *culture*?
   c. What about *emotions*?

B.