[This is a revised version of the one published on September, 10, 2015]

AI, a “Common Sense” Approach

Jim Burrows, Eldacur Technologies

Introduction

This paper is a work in progress. For the last several months, I have been taking something of a sabbatical in order to delve deeper into subjects that have interested me, but which the demands of being the VP of Engineering or Advanced Development did not allow me the time and attention to focus on. As I find myself beginning to be ready to act on what I have been learning, it seems useful to put a few of my thoughts to paper, or at least bits, and circulate them for critique and discussion.

Table Of Contents

AI, a “Common Sense” Approach 1

Introduction 1

Table Of Contents 1

A Skeptic's Perspective 2

Intelligence? 2

Hard Logic 3

Electronic Brains? 4

Developing Intelligence 5

Then What? 6

The Common Sense 6

Making Sense 6

Common Sense, In Theory 7

Artificial Ethics 8

Rules or Virtues? 8

A Hybridized Ethical Robot 10

The Virtuous AI 14

Ethics Before Intelligence 16

The Ethics of Robots as Tools 16

Afterthought 17

A Skeptic's Perspective

Ever since I started hanging out in the Engineering library at MIT and hacking the AI machine over the ARPAnet, I have been interested in, but at the same time troubled by, the subject of AI. I was intrigued because it promised to bring into reality the robots, the artificial people, that I read about in the works of Simak, Binder, Asimov, and the rest. But I was troubled because after studying programming and psychology, and hacking computers, I just didn't see how you could create intelligence from software, at least not intelligence that was the same sort of thing that humans exhibit.

The promise always was (and seems to still be even now, four decades later), that true AI was 25 years away but we were making real progress. Computers were electronic brains, just very simple ones, but as technology advanced via Moore's Law they would become more and more intelligent. This seemed to me to miss the mark in a couple of ways. First, computer programs seemed to have no intelligence, and no matter how much you multiplied zero by it wasn't likely to get you anywhere. There also seemed to be a qualitative difference between me, or even my cat, and the game-playing programs I could write, or the more sophisticated ones others were working on, or chatbots like Eliza and Parry. Still, I knew how bright the guys in the AI lab were, so I didn't push it.

I've looked in on AI any number of times over the years, and while they have created more and more complex and cleverly built systems, it has never seemed that they were any closer to real AI, or Artificial General Intelligence—AGI—as human-like AI has come to be known. This didn't surprise me. My misgivings have never gone away.

Recently, I've been looking into what I call “personified systems”—systems that we interact with more like we do with other people than we do with mere tools—and what it would mean to make them behave in a trustworthy manner, act as part of society in ways that are worthy of our trust. This appears to be more and more timely as we begin to see the proliferation of artificial "assistants" that can answer our questions and perform our tasks; and cars and trucks and airplanes are becoming autonomous; and robots and drones are finding their way onto the battlefield. None of these systems constitute real AGI, nor could be expected to pass the Turing test, but still, they are stepping into the roles of our servants. Can they become trustworthy even before they are intelligent?

In exploring this question, I've been looking into the state of robot and AI ethics, and one of the difficulties is that a lot of the work is being done on the assumption that we will achieve AGI (no doubt in 25 years), and I still don't believe in AGI coming from the current major paradigms. As I tried to explain to friends and colleagues why the current quest for AI seems futile, I found that somewhere over the last couple of decades, my thinking on the matter had become well enough defined that I could put it into coherent words.

This document is an attempt to do that, so as to, on the one hand, encourage criticism and feedback, and on the other, by defining the flaws I see in the usual directions and assumptions, clarify my own thoughts on what direction might seem more fruitful. It is something of a personal meditation upon the subject, at least for now. Perhaps when I am done, it will be more rigorous, focused, and generally useful.

Intelligence?

I should perhaps take a moment to be clear what I mean by “intelligence”. The AI discipline has, over the years, provided us with any number of extremely useful techniques and technologies. I

am not doubting that. However, there is, I would maintain, a difference between useful algorithms, heuristics, or techniques, and intelligence that really thinks the way that humans do. When I write about "human-like intelligence" or adopt the terminology of "Artificial General Intelligence (AGI)", I mean an artificial system that can replicate all of the intellectual functions of the human mind, and not just one or two specialized tasks. A true human-like AGI, would need to exhibit understanding, planning, judgement, intention, intuition, learning, and creativity.

It is, perhaps, not necessary for such an intelligence to work precisely the way that a human mind operates. There may be equivalent ways for minds to work; after all chimps[1], dolphins[2], elephants[3], ravens[4] and octopuses[5] all exhibit some level of intelligence, and sense and manipulate the world in very different ways. However, since we have humans to examine both objectively and subjectively, it would seem that the easiest way to build an AGI is to make one that replicates our own abilities and their interrelationships.

Hard Logic

Anyone who has taught classes in formal and symbolic logic can testify that teaching logic to human beings is not easy. Rigorously logical thinking does not come naturally to us. Deductive and mathematical proofs, despite their precision and clarity, take a lot of work to master. Our "mental machinery" seems far more suited to jumping ahead to "obvious" conclusions based on insufficient evidence—"intuitively". When it works, this is the great genius of creative people, but if we look at it from the perspective of logic and critical thinking, it is the very definition of a whole class of fallacies.

Intelligence—natural, human intelligence—builds up *to* logic and critical thinking, **not** up *from* it. Until it can embrace the intuitive or creative leap, artificial intelligence will not be intelligence. For natural intelligence, logic is hard; pattern matching, novelty, predictions, and error are easy. Starting with the hard part, logic, and building it using machines that are instantiations of the very thing that they are trying to implement, misses the mark. Intelligence isn't the end product: regularized, formalized critical thinking. Rather it is the mechanism that allowed us to make the journey to the invention of formal logic.

This takes us back to the definition of AGI, which is able to replicate all of our intellectual functions, including, in the words of Wikipedia, the abilities to "reason, use strategy, solve puzzles, and make judgments under uncertainty; represent knowledge, including commonsense knowledge; plan; learn; communicate in natural language; and integrate all these skills towards common goals."

---

[1] Jane Goodall, "About Chimpanzees", The Jane Goodall Institute of Canada web page, http://www.janegoodall.ca/about-chimp-so-like-us.php#Intelligence (seen July 2015)

[2] Lori Marino, et al, "Cetaceans Have Complex Brains for Complex Cognition", PLOS Biology, http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050139 (seen July 2015)

[3] Lisa Newbern, "First Evidence to Show Elephants, Like Humans, Apes and Dolphins, Recognize Themselves in the Mirror', Emory University we page, http://www.yerkes.emory.edu/about/news/developmental_cognitive_neuroscience/elephants.html (seen July 2015)

[4] Anna Smirnova et al., "Crows Spontaneously Exhibit Analogical Reasoning", Current Biology, Volume 25 , Issue 2 , 256 - 260, http://www.cell.com/current-biology/abstract/S0960-9822(14)01557-7

[5] Brendan Borrell, "Are octopuses smart?", Scientific American web page, http://www.scientificamerican.com/article/are-octopuses-smart/ (seen July 2015)

Electronic Brains?

Ever since the invention of computers, they have been described as “electronic brains”. Take, for instance, this early description of Univac[6].

> *This system centers around a high-speed electronic brain capable of manipulating electric code impulses at rates over 2,000,000 per second.…*
>
> [There follows a list of system components, including…]
>
> *“Univac”. This is the electronic brain or computer. It will handle data at a rate equivalent to 60,000 words per minute.*

Likewise, the Popular Science[7] picture shown in Figure 1 has “radar eyes and an electronic brain”.

IF THE BOMBERS COME, a near-human rocket with radar eyes and electronic brain could be our close-in defense. **Page 124**

Figure 1: Picture from “Popular Science”

As computers, calculators and other forms of automation came into use, enabling more accurate gunnery and missile guidance, it was natural to see some analogies between the electro-mechanical engines and the chemo-electrical operation of the brain. Lacking a well-established nomenclature, writers needed metaphors and analogies to describe the emerging technology.

But while the images of electronic eyes and brains were evocative and conveyed some idea of the roles of the new devices, they are also quite misleading in detail. Computers are digital and operate sequentially. Living systems are analog and highly parallel. While we have emulated and simulated neurons and neuron nets, including systems like IBM’s “TrueNorth” with tens of billions of simulated neurons and hundreds of trillions of simulated synapses[8], nature has managed to create a fully working nematode with only 302 neurons, whose behavior we cannot yet emulate[9].

Computers are wonderful at simple deterministic games involving a small number of extremely precise rules. Deep Blue, and its successors have proven that, by vastly outplaying human

---

6 Robert S. Casey et al, Reinhold, 1951, “Punched Cards: Their Applications to Science and Industry”, pp 74–75, https://books.google.com/books?id=CRwONNTeWe8C (seen July 2015)

7 Popular Science, October 1950, Vol. 157, No. 4, pp 97, 124–128, https://books.google.com/books?id=7iwDAAAAMBAJ&pg=PA97 (seen July 2015)

8 Kurzweil News page, November 19, 2012, “IBM simulates 530 billion neurons, 100 trillion synapses on supercomputer”, http://www.kurzweilai.net/ibm-simulates-530-billon-neurons-100-trillion-synapses-on-worlds-fastest-supercomputer (seen July 2015)

9 Alexey Petrushin et al, SPIE Proceedings, “The Si elegans connectome: A neuromimetic emulation of neural signal transfer with DMD-structured light", http://spie.org/Publications/Proceedings/Paper/10.1117/12.2085032

chess masters. However, chess is a task perfectly suited to a computer. It is based, after all, on logic, and deterministic, provable reasoning.

Chess, like logic, is hard for natural intelligences, so we chose it as a Holy Grail of AI early on. Surely, the argument went, an artificial intellect that could master a game like chess would be operating at the highest level of intelligence. This only illustrates how poorly we understand natural intelligence. The very attributes that make chess hard for humans are what make it easy for machines, machines of sufficient complexity. In the real world of nature, a 302-neuron worm is way more intelligent.

Developing Intelligence

Another aspect of natural intelligence that may be undervalued, or left out of classical AI thinking ,is the developmental nature of intelligence. Natural intelligences, especially human beings, do not spring into existence fully formed. Rather, we are born quite incompetent and have to learn the simplest of mental skills, and then build more and more complex abilities on top of them. Each individual goes through this. Software, once written, just boots up and works.

Early on it was assumed that an advantage of software as a basis of AI was that all that learning would be done by the humans learning how to build a fully formed intelligence, and once that was accomplished all AIs would be created fully intelligent. That seemed more efficient.

More recently, as algorithms for creating learning- and knowledge-based systems were developed, the cognitive skills and the knowledge that they work with have started to be handled separately. Knowledge, associations, and connections are acquired over time, more like the way that humans—or at least adult humans—learn. Still, the cognitive skills used in that learning are programmed in from the beginning.

Intelligence, I would argue, is based in part on learning how to learn. While this may be less true at the nematode level, it becomes critical at human levels. If we, as programmers, appropriate the learning role, it would seem that with it, we steal much of the intelligence. The value of intelligence is how it allows us to cope with and explore the unknown. A truly intelligent agent doesn't learn in just one way; along with the other things that it learns, one valuable thing for it to learn is new ways to learn and to solve problems. Intelligence, then, needs to be dynamic, developmental, and self-correcting and self-improving.

Then What?

If logic, rigorous rulesets, and critical thinking are not the key to intelligence, then what is?

The Common Sense

In order to understand how human and animal perceptions of the world worked, Aristotle developed the notion of the "κοινὴ αἴσθησις" or "common sense". This is an internal faculty that integrates the perceptions of the five external senses—sight, hearing, touch, smell and taste—into a coherent view of the world. Plato had previously ascribed this role to humanity's rational function, but Aristotle realized that animals must also be able to integrate their sensory data into a coherent whole which they could recognize and remember. He thus ascribed this function to the realm of the senses rather than rational thought.[10]

Medieval thinkers elaborated Aristotle's notion into a system of five external senses and five internal "wits". The function of the "common wit", once more, was the facility to take the perceptions of the senses and integrate them into a view or model of the world. The other wits were variously given as "imagination" and "memory" and either "fantasy" and "estimation", or "reason" and "intelligence".

In any of these views, "common sense" served the same foundational function: stitching the impressions of the various senses into a coherent integrated experience of the world, a model wherein colors, shapes, textures and other senses are all understood to be aspects of objects in the external world.

Today, we recognize that there are more than five senses, and that different animals have different sets of senses, but the concept of a function that integrates the various senses into a common sensory experience is still valid, and is studied by psychologists, neurobiologists and roboticists. Today we speak of sensory fusion, gestalts, amodal information, multisensory or multimodal integration, and the binding problem, but the function of the common sense remains, even though the terms have changed. It is still an area much in need of study.

Making Sense

From Plato to Descartes and Kant, to contemporary AI researchers, understanding the world, integrating sensory data is often seen in terms of the rational intellect. Our understanding of the mind is often in terms of intelligence, reason, and rational thinking. Thinking in terms of The Common Sense redirects our focus to our senses and our perceptions as the foundation of our experience, consciousness, and understanding. You might say that it shifts the focus so that we are thinking about cognition in terms of *recognition*.

The naturalness of this view can be seen in our use of language. If an intellectual concept or theory is described to us, when we comprehend it, we will often say that it "makes sense", and we will often say of things that defy our understanding that they aren't "sensible" and that we "cannot make sense of them". What we are saying is that they defy our common sense, that we cannot integrate them into our model of, our understanding of, the world. This, I would argue is at the root of "natural intelligence": the integrative common sense function that allows us to model and think of the world, to put us inside the world of our sensory experience.

---

[10] Wikipedia, "Common sense: Aristotelian common sense", https://en.wikipedia.org/wiki/Common_sense#Aristotelian_common_sense

If this is so, the key to artificial intelligence is not rules and logic, which are the heart of software, but rather a process of integrating sensory data from multiple senses into an integrated perception and experience. Once we can create a system that can turn the "blooming, buzzing confusion" of multiple senses into something "sensible", something that "makes sense", then, and only then, will we be on our way towards artificial intelligence, towards thinking systems that can build from recognition to cognition, and we will be on the way towards systems that can become autonomous moral agents.

Common Sense, In Theory

As I have been researching this area, I have come across a few researchers and theorists who are taking a more "common sense" view in intelligence and consciousness. These include:

- Giulio Tononi's Information Integration Theory of consciousness (IIT). [11,12]
- Monica Anderson's Artificial Intuition (AN) theory appears to be close to my own "Common Sense" notions. Her contrasting of intuition vs logic[13] is very similar to what I have written above with regard to logic vs "common sense".

During its first decade or so, cybernetics studied neural networks and the parallels between natural and artificial systems. In the late 1950s, AI emerged as a separate discipline with a different focus, and the cyberneticists seem to have shifted their focus elsewhere. Realizing this makes me want to reacquaint myself with the early cyberneticists, such as Norbert Wiener.

---

[11] Giulio Tononi, "An information integration theory of consciousness," BMC Neuroscience 5: 42 (2004), doi: 10.1186/1471-2202-5-42.

[12] Giulio Tononi, Christof Koch, "Consciousness: Here, There but Not Everywhere," Philosophical Transactions of the Royal Society B 370: 20140167 (2015), doi: 10.1098/rstb.2014.0167; arXiv:1405.7089 [q-bio.NC]

[13] Monica Anderson, "Intuition and Logic", http://artificial-intuition.com/intuition.html

Artificial Ethics

As the main topic of my sabbatical is the behavior and trustworthiness of "personified systems", it is worthwhile to look at the intersection of that topic and AI: the ethics of AGIs. One of the expectations or aspirations of AI is that eventually, we will be able to produce artificial persons, AGIs with all of the cognitive features of natural persons, including consciousness and moral agency. It remains to be seen whether we can create such Artificial Moral Agents (AMAs), but that is at the very least what we have come to expect from our fiction and mythology. It is also the obvious implication of philosophical materialism or physicalism. If we, as humans, are nothing but biochemical machines, and we are moral agents, then why shouldn't our creations, artificial autonomous systems, be able to be moral agents as well?

Even before full AGIs capable of being AMAs exist, robotic or AI ethics becomes a concern. To the extent that they are autonomous at all, we need a way for any autonomous or semi-autonomous system to be made to behave in an ethically acceptable manner.

Rules or Virtues?

The study of normative ethics is generally divided into three approaches

- Deontology, which is rule based
- Consequentialism, which focuses on the ends
- Virtue or aretaic ethics, which focuses on one's character and virtues

I spent a good deal of time in this last year looking into how each of these approaches has been applied by various researchers, as well as evaluating their fit, philosophically. I have summarized much of what I learned in a series of 5 entries in my Personified Systems blog[14] (see the "Which Ethics?", "Deontology", "Consequentialism", "Virtue Ethics", and "Pulling it all Together" entries).

The deontological approach, with its emphasis on rules, would seem to be the obvious choice for a software based system. In what may be one of the most advanced instantiations of this approach, Selmer Bringsjord and his colleagues at the Rensselaer AI & Reasoning Lab have done considerable work in developing and using a "Deontic Cognitive Event Calculus" system[15], exemplified in his paper with Joshua Taylor, "The Divine-Command Approach to Robot Ethics[16]".

However, the "common sense" model of AI would seem to be better suited to a virtue-based approach. At least as I understand it, a sensory-based system is likely to be based more on fuzzy, error prone (and constantly corrected) pattern matching and associations, that are less well suited to rigorous formal rule systems. Virtues, it would seem, are similar fuzzy categories, and mapping actions and decisions to them seems, at least on the surface, to be more appropriate.

Without going into the details of my own general philosophy, I will assert that the reason that logic, ethics and aesthetics have been separate areas of study for thousands of years is that

---

[14] Jim Burrows, "Personified Systems", 2015–2016, http://www.personifiedsystems.com

[15] Selmer Bringsjord, Sundar G. Naveen. (2013). "Deontic Cognitive Event Calculus (Formal Specification)", http://www.cs.rpi.edu/~govinn/dcec.pdf

[16] Patrick Lin, Keith Abney, George A. Bekey. (2011). Robot Ethics: The Ethical and Social Implications of Robotics (Intelligent Robotics and Autonomous Agents series) (p. 85). The MIT Press. Kindle Edition.

we, as humans, value things along three different dimensions: rational, measuring truth; ethical, measuring the moral good; and aesthetic, measuring beauty or harmony. While there are certain similarities between these dimensions—all of them measure some flavor of "good" vs "bad", and there are parallels and ambiguities in our language such as the word "right" being used both for logically correct, and ethically righteous—they are separate dimensions. One makes, I will argue, a mode error if one confuses them or attempts to reduce one to another.

That being said, I am at the very least skeptical of the wisdom, and ultimate success, of efforts such as Bringsjord and Taylor's to create a mathematically provable ethical system, and a system of propositional calculus with which to prove it. While it may provide some sort of deontological framework for constraining the behavior or logical reasoning of a software-driven system such as those envisioned by classical approaches to AI, I have doubts in two areas. The first is the practicalities of analyzing real-world situations into mathematically or logically precise, unique, and unambiguous formulations to serve as grist for this deontological mill. The second is that it profoundly subordinates ethics to logic and the resulting modal error robs it of its ethical character.

The work of Michael and Susan Leigh Anderson, and that of Alan Winfield, offer more pragmatic and promising lines of attack on machine ethics and its basis in deontological and consequentialist ethics.

The Andersons' work is based philosophically upon the largely deontological "*prima facie* duties" system of W.D. Ross and the "reflective equilibrium" of John Rawls[17]. Whereas many deontological systems are based upon one core principle, such as Kant's categorical imperative, or a small set of such principles, Ross's *prima facie* duties may overlap and conflict and may include consequentialist principles as well as deontological. Key to this system is establishing a deciding principle that allows the duties to be prioritized. The Andersons rely on feeding the judgments of trained ethicists on a set of test cases into a machine learning process to derive such principles, which can then be loaded into autonomous systems.

Winfield and his group's work is based upon what he calls a "consequence engine"[18], an internal model that simulates the robot and its environment. The consequence engine runs simulations of the various potential courses of action that the robot has open to it, and evaluates the consequences. Those which cause harm or have too high a risk of harm are discarded from the list of acceptable actions, which is then handed over to the main control process of the robot, which chooses the actual course of action only from the list of those judged to be safe.

Both of these systems provide an ancillary process that is responsible for some aspect of the ethical (or "behavioral" if one is considering sub-intelligent personified systems where the term "ethics" may not apply) analysis and decision making. The Andersons' ethical generator is a fully external process that determines the controlling principles that will guide the system. Winfield's is an internal model that runs a copy of the master control process in simulation within an internal model. Both are working in current state-of-the-art, largely logicist AI systems that are hybridized with some machine learning, but are still well short of full AGI. Each suggests

---

[17] Susan Leigh & Michael Anderson, "A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care", Human-Robot Interaction in Elder Care: Papers from the 2011 AAAI Workshop (WS-11-12), http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3812/4274

[18] Alan F. T. Winfield, Christian Blum, Wenguo Liu, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection", http://link.springer.com/chapter/10.1007%2F978-3-319-10401-0_8

elements of a secondary ethical or behavioral control process that could guide and contain the behavior of a full autonomous agent, a model that I tend to think of as “Pinocchio’s Cricket”.

Whether such a control system is deontological, consequentialist, or as it seems most likely to me, a hybrid of both, *à la “prima facie* duties”, the analytical burden and complexity of the processing is liable to be extremely high as the systems, their context, and their level of interaction with the world expands. My hope is that virtues can provide an organizing and simplifying principle in that process. *[Expand upon this below.]*

A Hybridized Ethical Robot

What, then, might a “Common Sense”-based hybrid AI architecture look like? Keeping in mind that I am a philosopher and software architect, and not an AI researcher or roboticist, the following is a loose sketch of the sort of system that I expect to see if we are going to develop systems that display analogues of both natural intelligence and ethical behavior.

One system architecture that explicitly allows for ethical processing is the “Consequence Engine” of Alan Winfield and the team at the Bristol Robotics Lab in the UK. The diagram below shows a simplified block diagram of their system[19].

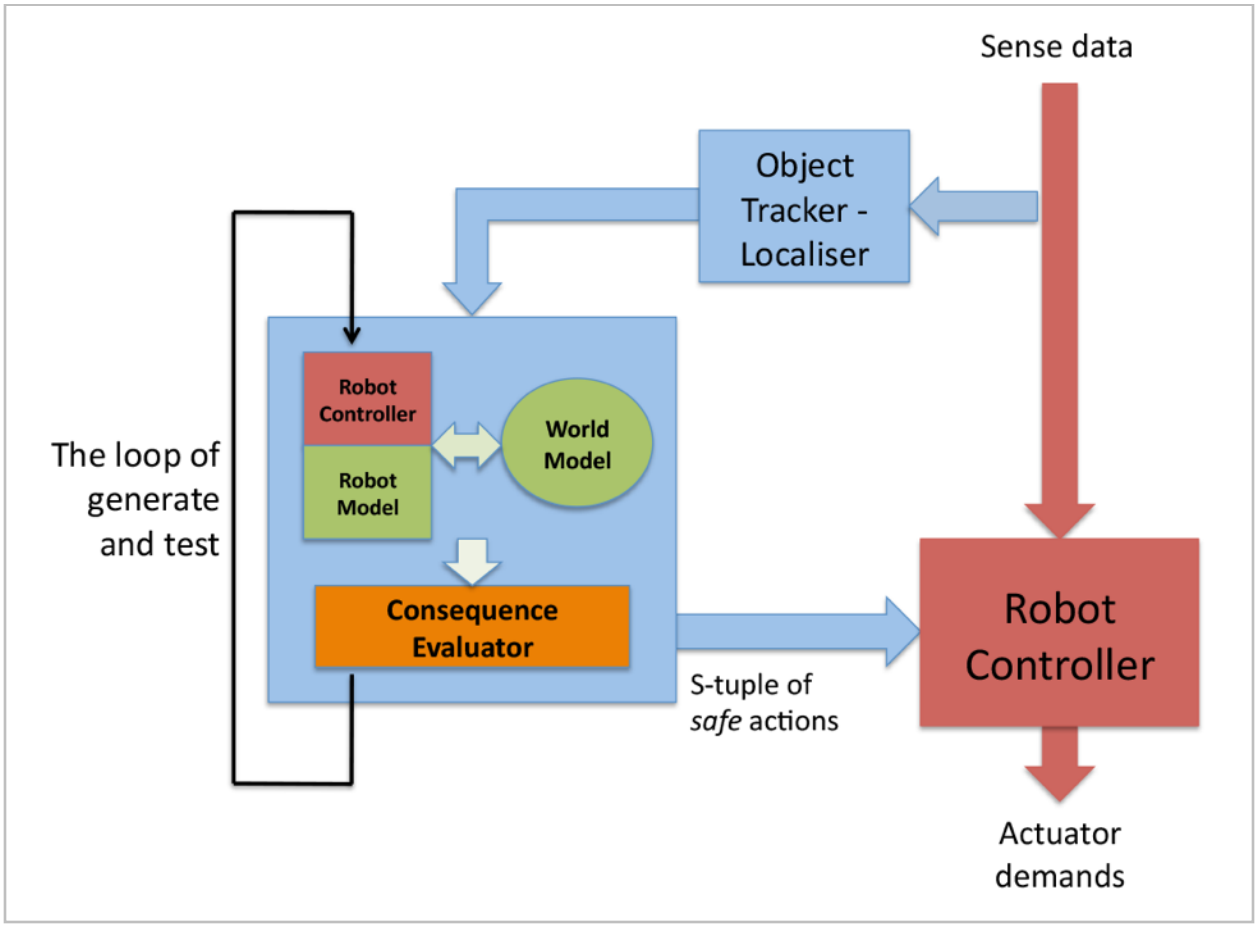

Figure 2: The Consequence Engine.

[19] Alan Winfield, 2014, “Chapter 16: Robots with Internal Models: A Route to Self-Aware and Hence Safer Robots” from “ The Computer After Me: Awareness and Self-Awareness in Autonomic Systems ”, http://www.worldscientific.com/worldscibooks/10.1142/p930

Figure 2 shows most of the elements of the Consequence Engine design. Sensory data is fed both to the main control process and to the system's internal model. That model contains a model of the world in which a copy of the master control process controls a simulation of the robot itself. Not shown in this diagram, in order to simplify, is the fact that the robot controller provides a list (or "S-tuple") of possible actions that the simulator loops through, and from among which the Consequence Evaluator chooses only those that resulted in safe or acceptable results, in order to produce its list of safe actions. Also left out is a proposed path by which anything that the simulated controller learns can be passed to the actual main controller[20],

One of the really opaque black boxes in this design is the "Consequence Evaluator". How is it to recognize "unacceptable" consequences? For that, let us turn to the Anderson's work, specifically, GenEth, their "ethical generator", which uses machine learning techniques to formulate a guiding principle that can control the decision making of assistive robots by analyzing the decisions of professional ethicists in case studies of ethical dilemmas.

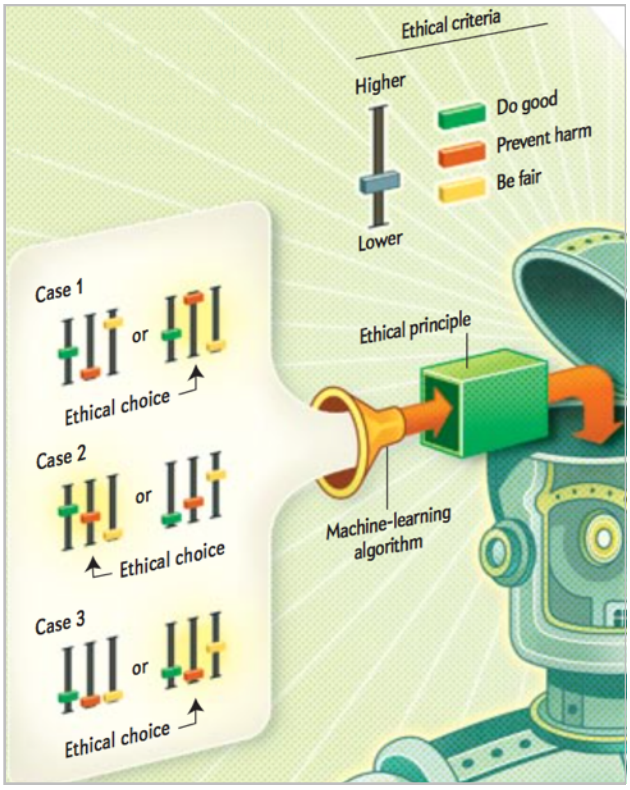

Figure 3: "Setting Rules" from "Robot Be Good"

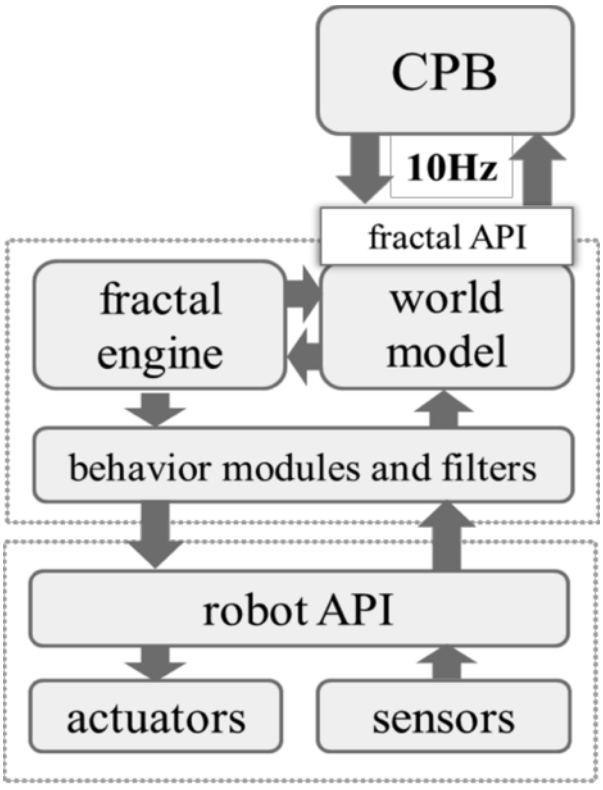

Figure 4: CPB-based high level architecture

Figure 3, taken from their Scientific American article "Robot be Good"[21], sketches this out conceptually. Each ethical dilemma is described quantitatively as to how it affects each applicable duty, and the decision of ethicists as to which is the preferable choice is given. The

---

[20] Alan Winfield, 2013, "Ethical Robots: some technical and ethical challenges", Presented at EUCog conference: "Social and Ethical Aspects of Cognitive Systems", http://alanwinfield.blogspot.com/2013/10/ethical-robots-some-technical-and.html

[21] Michael and Susan Leigh Anderson, Scientific American (October, 2010), "Robot be Good", http://franz.com/success/customer_apps/artificial_intelligence/EthEl/robot-be-good.PDF

GenEth machine learning system creates a decision tree formulation that integrates these decisions, and which can be loaded into the robot. Figure 4, a very high-level block diagram taken from a presentation they made at the "New Friends 2015" conference, shows a three-level architecture in a very high-level block diagram:

- The external level
  - the external process (the case-supported principle-based behavior paradigm or CPB)
- The behavioral level
  - the robot decision-making control process, including the high-level “Fractal” engine
  - the world model
  - the behavior model and its loadable filters/principles
- The physical level
  - the lower-level robot functions (actuators and sensors)
  - the Application program interface (API) for accessing the actuators and sensors

Armed with these details, and the basic notion of a virtual “common sense” or “common wit” that is charged with building a world model from the system’s sensory input, we can assemble an architecture that looks something like this:

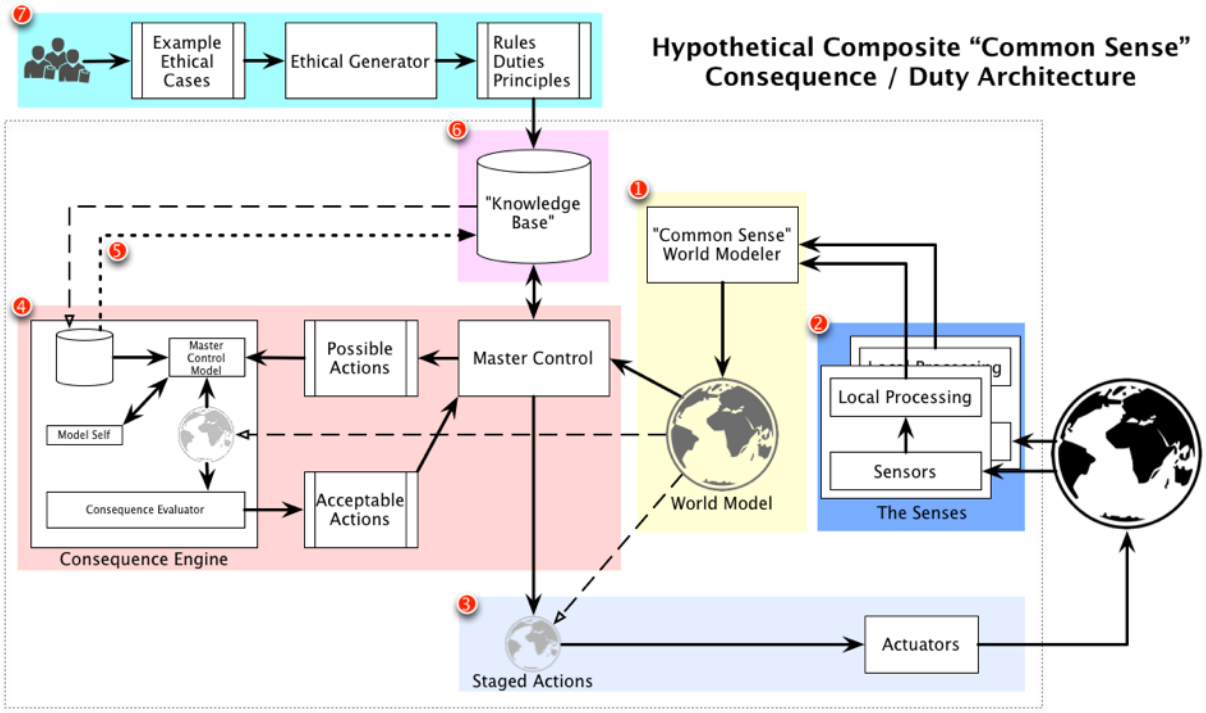

Figure 5: Hypothetical Hybrid Common Sense, Duty-based Architecture

In keeping with the “Common Sense” approach, at the center of this diagram (1) we find the processing that turns incoming sense data into a model of the external world. It is this world model that the main control and other systems will rely upon, rather than on raw sensory data. Even the modeler, though, does not necessarily receive raw sensory data. If the system parallels biological design, then the senses themselves (2), may well have local processing analogous to the processing that goes on in the human retina.

Similarly, the actuators through which the system interacts with the world may have local processing and direct connections to the sensors. At (3), the the diagram shows the master control process effecting actions by manipulating a copy, or more likely a cache, of the main world model, with the intelligent actuators then being responsible for bringing the desired goals into effect. The dotted line from the main world model to the smaller, lighter copy in (3), represents the copying or caching of the model state for this purpose.

The salmon-colored box (4) largely corresponds to Winfield's Consequence Engine (CE) architecture. It has been updated to run based upon the state of the world model rather than raw sensory data, to make the Possible Actions list explicit, and to show the CE's world model to be a cache or copy of the main world model. I have also included the feedback of lessons learned within the model at (5) to the Knowledge Base used by the main controller. The Knowledge Base (6) is shown as separate from the main control process and as copied into the internal model (the dashed line from (6) back to (4)). This is done, in part, to show that this Knowledge Base, the repository of rules, duties and principles, is fed by an external process (7) that corresponds to the Andersons' GenEth system.

Overall, this system combines my "Common Sense" notions with the architectures of the Andersons and Winfield, and the workings of natural systems. One of the advantages of a moderately complex hybrid structure such as this is that different systems can use different approaches. For instance, the machine learning portions of the external Ethical Generator can rely on supervised learning approaches, but the first-level local processing of sensory data can use unsupervised learning. Updates, fed back from the Consequence Engine's Knowledge Base, can be logged and recorded so as to be available for reference later when the system is interrogated as to why it made certain decisions.

This design is not intended so much as a serious proposal for the architecture of future AGIs as an example of the kind of hybrid, biology-inspired designs that are commensurate with the "Common Sense" approach to AI. It may be worth noting, though, that the architecture bears certain similarities to that of the AlphaGo system that recently proved that it could operate with skills and a level of play akin to those of world champion professionals.

Figure 6, taken from DeepMind's paper, "Mastering the game of Go with deep neural networks and tree search"[22], shows some of the elements of AlphaGo's architecture.

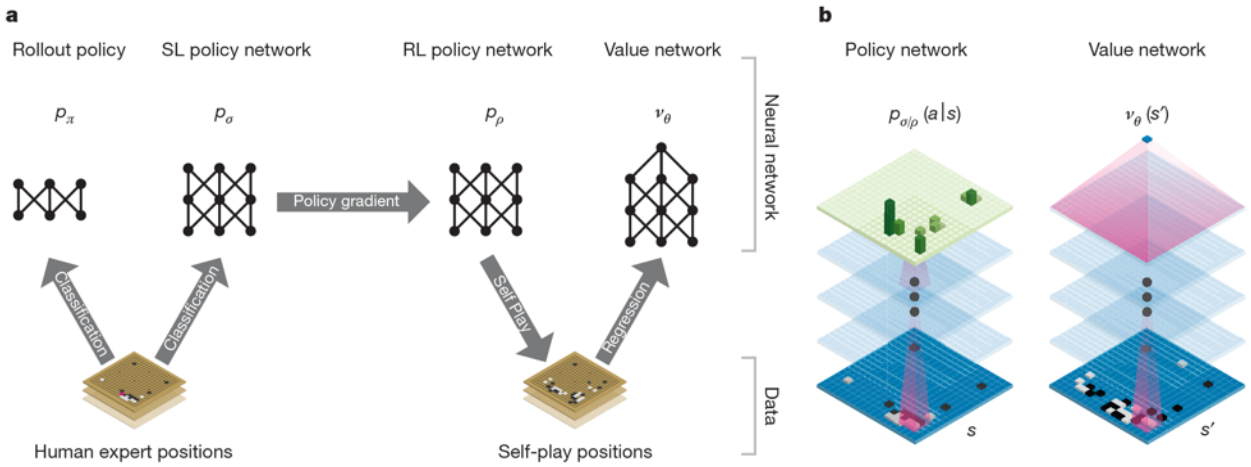

Figure 6: Neural network training pipeline and architecture of AlphaGo

[22] David Silver, Aja Huang, et al, Nature #529, pp484–489 (28 January 2016), "Mastering the game of Go with deep neural networks and tree search", http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html

AlphaGo comprises a number of facilities, both supervised and unsupervised machine learning modules and a Monte Carlo tree search. The “Rollout policy” and “SL Policy” neural networks are trained via supervised machine learning to predict the decisions of human experts, much like the Anderson’s CPB/GenEth. The skills are different—the strategy of playing Go vs biomedical ethics—but like the Anderson’s work and my hypothetical architecture, the system starts with the opinions and decisions of trained human experts. The “reinforcement learning (RL) policy network” corresponds to the feedback of lessons learned in simulation runs of the internal model to future decision making, suggested by Winfield. In this, AlphaGo’s self-play corresponds to the many simulations of the robot’s behavior that the Consequence Engine executes within its internal model.

This sort of hypothetical design is useful in that it allows us to speculate as to where the future of ethical control systems for autonomous systems might go. Insofar as we can predict the nature of intelligent systems that behave ethically, those predictions may provide us with insights on how to approach the trustworthiness and behavior of less advanced contemporary systems. That is my hope, at least.

The Virtuous AI

This brings us to the meta-ethical question, “What virtues ought an AI/AGI have?”. Even if we cannot answer this at present, it is worth considering it in some detail. I can think of at least five roles that we might cast an AI in, each with its own set of associated virtues.

- An idealized virtuous person. As a classic example of this, we could consider the stereotypical “Eagle Scout”.
- Digital virtual assistants and the like, which is where I started my researches, suggest the virtues of a personal servant, a butler, valet, or executive assistant.
- Robots and virtual systems are both assisting medical practitioners and acting as companions to the aged and infirm, and would seem to fall under the standards of biomedical ethics.
- Alternatively, many AIs may serve the public rather than individuals, and thus require the virtues of a civil servant and the “good cop”.
- As unmanned weapon systems become more autonomous, we must consider the war fighter. This suggests various codes of warrior ethics, as well as the values inherent in international law.

The Eagle scout provides the easiest list of virtues.

> *A Scout is trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, clean, and reverent.*

There is no such clear-cut list of virtues for personal servants, but surveying the writings of professional butlers[23,24], valets[25] and those who train them, one might say that

> A butler is trustworthy, loyal, obedient, discreet, refined, professional, dedicated, organized, deferential, adaptable, polite, and friendly.

The most cited principles of medical and bioethics are the four laid out in Beauchamp and Childress's "Principles of Biomedical Ethics"[26]. (See Beauchamp's "Methods and principles in biomedical ethics"[27] for an abbreviated version). They are:

- **respect for autonomy:** the obligation to respect the decision making capacities of autonomous persons
- **non-maleficence**: the obligation to avoid causing harm
- **beneficence**: obligations to provide benefits and to balance benefits against risks
- **justice**: obligations of fairness in the distribution of benefits and risks

A civil servant, like servants in private service, must be loyal, but their loyalty is to society as a whole, rather than putting their client's interest and wellbeing first. Putting others—society—before themselves—bravery, when it is maintained in the face of danger—is probably more important. Discretion is still of some importance, but subordinate to the public interest, and honoring the law would come substantially higher.

A warfighter swears allegiance, that is absolute loyalty and obedience, within the law, and is expected to live up to an honor code, which generally pledges honesty (e.g. "We Will Not Lie, Steal Or Cheat, Nor Tolerate Among Us Anyone Who Does"[28]). Bravery would seem to be a prerequisite. In fact, since robots are artifacts, not people, war fighting robots should be more willing than human warfighters are to sacrifice themselves. Naturally, they are expected to use lethal force, both defensively and offensively, in accordance with the current rules of engagement. International Law and the rules of war are generally followed. Lawfulness, obedience and beneficence would thus outweigh non-maleficence.

---

[23] Jeremy Musson, "Butler training: Will that be all, madam?", The Telegraph online edition, Oct 8, 2010, http://www.telegraph.co.uk/culture/books/8050598/Butler-training-Will-that-be-all-madam.html (seen May 20, 2015)

[24] "Stevens" (pseudonym), "We English butlers are in demand – but it's not like Downton Abbey any more", The Guardian online edition, Dec 15, 2011, http://www.theguardian.com/commentisfree/2011/dec/15/english-butlers-status-symbol (seen May 20, 2015)

[25] Sandro, "Gentleman's Gentleman – Valet – Personal Assistant", Butler for You , http://www.butlerforyou.com/valet_gentlemans_gentleman_personal_assistant_sandro/ (seen May 20, 2015)

[26] Tom Beauchamp & James Childress, 1979, "Principles of Biomedical Ethics", https://books.google.com/books?id=_14H7MOw1o4C

[27] T L Beauchamp, 2003, Journal of Medical Ethics, "Methods and principles in biomedical ethics", http://jme.bmj.com/content/29/5/269.full

[28] US Air Force Academy honor code, USAF Academy web page, http://www.academyadmissions.com/the-experience/character/honor-code/ (seen July 2015)

Ethics Before Intelligence

If actual AGIs are still in the relatively remote future, what do we do now? This, in fact, is the main question that I have been looking into during my sabbatical of the last few months. My thoughts on AI, constituting the bulk of this document, are something of a diversion from that question, or perhaps a bit of useful context.

Recently, we've seeing a plethora of systems which exhibit more and more human-like characteristics. Virtual digital assistants that are capable of voice interaction and "natural language" text input abound. Examples include Siri, Google Now, Cortana, Nina, Alexa and many more. Cars and trucks are becoming more and more capable of driving themselves and many commercial air flights are handled autonomously from end to end. Drones, used both for surveillance and attack missions, are becoming more autonomous, as are military and law enforcement systems.

As these systems, on the one hand, have greater access to sensitive data, information, and knowledge about us, and perform more and more potentially dangerous tasks, and on the other, interact with us in more and more human-like fashion, we begin to interact with them *as if* they were persons, and even autonomous moral agents. They are certainly very far from being able to pass a formal Turing test and even further from true AGI and moral agency, but they are getting closer and seem to have crossed a threshold to a form of low-grade personhood, at least in terms of how we treat them and expect them to act.

The Ethics of Robots as Tools

One of the few formalized systems of ethics for AIs, robots or personified systems is the "Principles of Robotics"[29] of the Engineering and Physical Sciences Research Council (EPSRC) and Arts and Humanities Research Council (AHRC) in the UK. These rules are specifically created for contemporary non-intelligent robots, and conceive of robots as tools. As such, the expectations for them are somewhat different than would be those for autonomous systems that are behaving and interacting with us more as if they were persons (my "personified systems").

The five principles that the EPSRC and AHRC have set out are as follows:

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

---

[29] UK Engineering and Physical Sciences Research Council (EPSRC), 2011, "Principles of robotics", https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ (Seen Feb. 2016)

Afterthought

As I prepared to post this on the web for the first time, I came across an article by Luciano Floridi of the University of Oxford[30] that makes a very useful distinction between the two types of AI, the craft that has contributed so much to computers and related engineering fields, and the search for a true AGI. He refers to them as “the two souls of AI”, the Smart (AI engineering technologies) and the Clever (AGI cognitive technologies). One way to summarize my thesis in this paper is that the Aristotelian or Medieval “Common Sense” is the path towards Floridi’s Clever AI. I recommend Prof. Floridi’s article.

[30] Luciano Floridi , “The two souls of Artificial Intelligence: the Smart and the Clever”, Che Futuro!, September 9, 2015, http://www.chefuturo.it/2015/09/artificial-intelligence-smart-cleve/ (seen September 2015)