THE VALUE ALIGNMENT PROBLEM © Dan Bruiger August, 2021 dbruiger [at] telus.net

Introduction

The seminal work of Nick Bostrom [2014] galvanized thought and discussion about AI safety and the Value Alignment Problem, which presupposes that artificial general intelligence (AGI) is desirable and perhaps inevitable. It addresses the question of how to align autonomous AI entities with human values, goals and purposes. It is motivated by the fear that such entities, if super-humanly intelligent, could evade human control and come to threaten, dominate, or even supersede humanity.

However feasible or not, the project to imbue AI with human purposes or values is hampered by a) the difficulty of formulating them unambiguously; and b) inconsistency within the human community and over time. The VAP is a worthwhile exercise if for no other reason than that it holds up a mirror to understand the problem of aligning values among and within human beings.

AI maps human intelligence onto machine intelligence: if humans have objectives and pursue them, then so machines should have objectives and pursue them [Russell, p176]. This conflation follows historically from Descartes' belief that the essence of the human being—as opposed to the animal being—lies in the capacity for reason. However, the essence of goal-seeking for all agents, as for humans, lies in desire and feeling, not reason.

The VAP, as usually conceived, is one side of the more general issue of *mutual control* between agonistic agents. The *inverse VAP* is the problem (for humans) that AI might align human values with goals that appear arbitrary or adverse to people. (The mundane power of advertising and propaganda come to mind.) More generally, the control problem involved in the VAP is to avoid unintended consequences: to get the system to "do what I mean" rather than to literally do what it is told to do.

While some transhumanists might consider a superintelligent takeover desirable, probably most people would not. The prospect raises many questions, the first being whether AGI is inevitable or even desirable. The VAP invites clarification of key concepts such as 'agent', 'general intelligence', 'value', 'goal' ('final' and 'instrumental'), 'alignment', etc. One aim here will be to explore tacit assumptions and possible inconsistencies lurking behind such terms, many of which are somewhat recklessly imported from ordinary language. As is common practice in the field, the literature is full of loose usages in the domain of common human experience, glibly transferred to AI. Examples of typical "as if" language occur when a machine is said to *reason* or *know,* to have *incentives, desires* or *motivations*, etc. Perhaps these can be excused as metaphor or as applications of Dennett's intentional stance. Yet, they may mask genuine confusions and distinctions that are important to clarify.

Bostrom [p185] puts his finger on a key question to address: Could one have a general intelligence that is not an agent? This points immediately to the need to clarify what general intelligence is and what constitutes an agent; and then to questions such as *whose* values are to be aligned and the feasibility and conditions for "friendliness." Related further questions concern Bostrom's 'orthogonality thesis' and the correlated 'instrumental convergence thesis'.

However, a different sort of question concerns various levels of human motivation in relation to the promise of AGI. Why bother at all to build a superintelligence that has a will of its own? Likely it is

assumed that there is some advantage to consolidating diverse functions in a single, autonomous, self-modifying agent; but this would also increase the challenge of aligning AI behavior with human goals [Drexler Sec 13.5]. Especially with the project to create an artificial *agent*, effort and responsibility are transferred from the diverse tasks of individual software development to the mythical catch-all task of value alignment. The user-interface is simplified, but the entity created is more complex and along with it the programmer's responsibility. The VAP transfers to AGI the age-old dilemma of aligning values among human beings. It proposes to substitute a technical project for a social one. Instead of choosing to develop autonomous artificial agents, we could choose instead to more deliberately embrace the social/political challenge to coordinate and harmonize *people*.

Instead of consolidating skill in an *agent* (which acts on its *own* behalf), it would be safer to create ad hoc task-oriented software *tools* that do what they are programmed to do because their capacity to self-improve has been deliberately limited. This is the approach advocated by Drexler. An alternative path, advocated by Russell, would be to build uncertainty into AI systems, so that they hesitate before acting in ways adverse to human purposes. Thus, they would be obliged to consult with humans for guidance.

Intelligence

There have been many definitions of 'intelligence', a nebulous term that can refer to a *property* of an organism or system, or an *entity* that is supposed to be intelligent. The latter is the implicit sense when referring to *an* AI. In the case of GOFAI, clearly the intelligence involved is that of its programmers. The issue is less clear when the AI is self-programming or self-modifying. In the case of an agent, its intelligence is its own.

Intelligence is usually measured through accomplished tasks. It is thus relative to what the testers value and propose as tasks. Since we consider ourselves the most intelligent agent on the planet, the standard for intelligence is implicitly human. It seems useful to generalize the idea of intelligence in such a way as to liberate it from the human standard, which is not only anthropocentric but also represents a ceiling. The problem, of course, is that we must approach this generalized notion from the viewpoint of our own human intelligence. If superintelligence (SI) can transcend the human ceiling, *we* nevertheless remain limited by it; we may not be able to fathom, let alone measure, superintelligent reasoning.

On the other hand, ability to *reason* may not be the right basis for an intelligent concept of intelligence. Moreover, intelligence is not a *property* residing within an individual so much as a response and *relationship* to the world, which consists in large part of other intelligent beings. The idea of "pure intelligence" derives from abstracting certain abilities, through performance tests of various sorts, from their social and real-world contexts. Such abilities are then reified as internal powers of the agent concerned. Yet, what it means *interactively* for A to have greater intelligence than B is that the power of A over B exceeds the power of B over A.

An assumption implicated in the notion of *general* intelligence is that intelligence is a capacity independent of specific goals. This is not true of organisms, even of human beings. Such independence (orthogonality) is a human *ideal*, extrapolating from the *relative* freedom that humans enjoy to set arbitrary goals that have little to do with biological goals or limitations. The ideal is to create an entity that has absolute (or at least greater) freedom, beyond the limits

imposed by human embodiment. However, the intelligence we know through experience with people and other creatures is a property of living things. The *artifacts* we describe as intelligent reflect *human* intelligence. The notion of general or *universal* intelligence [Legg & Hutter, p4] raises the question: To what extent and in what ways is intelligence tied to biology? To what degree and under what conditions (if at all) can intelligence be abstracted from its organic models and exemplars, to serve as a logically coherent stand-alone concept that can inform the notion of artificial general intelligence (AGI)?

Certainly, the performance of organisms depends on cognitive abilities, which depend on sensory equipment and specific adaptations to environment, in which motor interaction plays a crucial role. Moreover, meaning and motivation depend on having a stake in survival and reproduction. (Emotion is not *ir*rational, but an alternative, "first-aid," system of response.) Abstractly, one could say that the natural purpose of intelligence is to reduce uncertainty—e.g., by finding an algorithm that does not already exist or a routine to replace conscious deliberation.

Like an organism, an AI learns through an interactive cycle of perception/action. But the training data-set in organisms is not supplied by humans but by the organism itself and its environment. "Correctness" of the learning is regulated by natural selection, not through rewards given by an external judge or trainer. Without an existential investment in value based on survival, adjusting reality to fit the goal would be just as valid as adjusting the goal to fit reality. Bostrom [ibid, p35] recognizes that "there is no reason to expect a generic AI to be motivated by love or hate or pride or any other such common human sentiment…" But why expect any motivation at all if 'generic' implicitly means *disembodied*?

Agency

In one sense, an *agent* is anything that does something. Its classical counterpart is a *patient*, meaning the passive recipient of an action, the effect of a cause. Each link of a causal chain plays both roles in the transmission of a force. These are arbitrary demarcations within a system, as is the definition or boundary of the system itself. Ultimately, there must be a first cause that initiates the causal sequence from outside the system. And ultimately the causal system concerned is unbounded: the universe as a whole. While God was traditionally needed as first cause—as the only plausible force outside this closed system—alternatively there is no linear "first" cause in a system of circular causation.

In a quite different and more restricted sense, an agent *is* a first cause. This is the sense of agency we normally and personally develop as human beings. It is learned and modelled on our ability to literally move our own limbs and thereby move other objects in space. (We do not experience such willing as a causal sequence within us, initiated from without, but as a process that originates with us.) In other words, we experience ourselves and others (including animals and various spiritual entities) as autonomous free agents, independent of any causal chain. At one time in human thinking (and still in some cultures), potentially everything could be an agent in that sense. There would have been an adaptive advantage for early humans to be alert to potentially dangerous agents, whether animal predators or other humans from rival groups [Atran, p77].

The category of *patient* (that is, inert matter) is a relatively recent development. Perhaps beginning with the ancient Greeks, it was inherited by the Semitic religions that served as matrix for the Scientific

Revolution. The world for us now remains ontologically divided into objects and subjects (agents)—passive matter and active mind. Like God, the mind of the observer (e.g., scientist) stands outside the system observed. This dualism translates into the two extremes of autonomous AGI, on the one hand, and hand-written computer code, on the other. It does not accommodate well the nebulous range of possibilities between.

The only true agents we know are organisms, especially other people. Such an agent is an *autopoietic* system [Maturana & Varela]: a system that is self-producing, self-maintaining, and self-defining. (In the case of life, it is also necessarily self-reproducing, since life-forms evolved through selection over generations.) Agents are also necessarily *embodied*—meaning not just physically instantiated but having a mutual, though dependent, interactive relationship with the environment.[1]

For an agent (i.e., a truly autonomous, autopoietic system), the only "final" or "ultimate" goal is its own existence. In that sense, the notion of an alterable or arbitrary *final* goal seems nonsensical. For an agent, self-preservation is not an instrumental goal, at the bottom of a hierarchy of sub-goals. Rather it is the unique final goal, at the apex of any hierarchy. An AI that is an autopoietic system will normally control its own inputs, gather its own information, define and negotiate its own relationships with the world, and may attempt to control the forces or agents attempting to control it. Yet, that is but one extreme (represented by AGI) of a range of possibilities. At the other extreme are conventional AI tools that are not agents.

To follow or obey a command is a different action for an AI *agent* than it is for an AI *tool*. An agent decides for itself how to interpret the input (in the light of its own needs and goals) and whether and how to respond to the command. Just as it is for humans, a *goal* for an AI should be distinguished from *commands* given to it. A machine "obeys" a command automatically, with no intervening will and no goals of its own. An agent may or may not embrace the programmer's goal as its own, weighed against the backdrop of the agent's own priorities.

Can a superintelligent AI that is not an agent pose a serious threat? Under what conditions is an agent a *mind* (and what exactly constitutes mind)? Can a mind be considered a formalizable system (software)? If so, what are its limitations in the light of Gödel's paradoxes? Can the parts of an agent be agents in their own right?[2] Since embodiment is a necessary aspect of agency, biology may have much to teach AI researchers. On the other hand, since biologists have inherited the same mechanistic worldview as AI researchers, they might learn from questions that AI brings to the fore.[3]

Goals, orthogonality, and instrumental convergence

The weak point of defining intelligence as the ability to accomplish goals [Tegmark] is that 'goal' is undefined and it is unclear *whose* goals are concerned. Concepts of goals, motivations, mind and agency

---

[1] As Maturana & Varela point out, 'environment' is a concept in the human cognitive domain; an organism may have a different representation of its surroundings or none at all.

[2] Tissues have latent holographic capabilities, which are constrained by their role in the whole organism.

[3] E.g., see [Ebrahimkhani and Levin]: "The existence of designed biobots pushes us to ask what exactly is a machine, what features might life have that are permanently… beyond the reach of engineers if they have access to the same ingredients and evolutionary methods used by natural life, and what precisely would make something not a machine?... As biobots and other forms of artificial life have the benefit of both design and evolutionary dynamics… there is no principled reason why future versions could not enjoy the same agency that… evolved lineages do."

derive from human experience of dealing with other apparent agents. As self-conscious beings capable of reason and abstraction (indeed, having invented them), we conceive and pursue goals that appear to be independent of our final goal as organisms—even sometimes contrary to it. This gives the impression that we can arbitrarily adopt any goal to be pursued by any means (orthogonality). While true in principle as an ideal, it is not so in fact. Like other organisms, humans remain motivated very much by considerations grounded in their embodied existence. We might expect the same from AGI. On the other hand, *non*-autopoietic systems have no motivation at all, no goals of their own. They could not, as suggested by Bostrom, bring to bear self-preservation as an instrumental sub-goal in service of some (arbitrary or humanly specified) "final" goal such as winning at chess—unless they were specifically instructed to do so. Deep Blue *could* be given access to unlimited resources and commanded to enhance its ability to play chess by taking over the Internet or manufacturing an indefinite number of copies of itself to be hooked up in parallel. And that could spell an environmental and existential catastrophe. But it would be an insane *human* decision, not Deep Blue's.

The dilemma of imparting *stable* goals is that a self-modifying AI may modify what the programmer has initially specified as its goal. It might do this because of a *meta-goal*, deliberately installed at the outset by human programmers to optimize some function. But if we suppose that it "naturally" modifies its goal through "reasoned understanding," then we must presume an agent with its own purposes. In that case, the human goals concerned must be *negotiated* with the AI, as they would be with other humans or animal agents. Moreover, an AI that is not an agent cannot "care" about anything, including its own effectiveness. Stability for an agent is not an instrumental goal to ensure its final goal, but an inherent aspect of autopoiesis.[4]

Bostrom offers no absolute distinction between final and instrumental goals (anything can be either, in a hierarchical relationship). Instead, he promotes the lack of this distinction as a general principle.[5] The orthogonality thesis abstracts the human situation of a self (executive function) that happens to be *relatively* detached from the biological goals of the body as an organism. This ideal does not characterize living things, however, and would not characterize artificial agents by default.

The correlated instrumental convergence thesis tacitly *presumes* agent status while (by the orthogonality thesis) it disclaims any dependence on it. The properties that are framed as "instrumental" include self-preservation, goal stability, self-enhancement, expanding real-world resources and power. All of these, however, are rather aspects of autopoiesis.

Unless an AI happens to be an autopoietic system, its goals could be only those of its programmers, its only actions a function of its program. The tricky part is when that program is unknown, for example because it is a neural network or some other self-organizing or self-improving system. To presume to control the evolution of such a system (a black box) by setting its "initial conditions" is problematic if not paradoxical, since what the black box contains cannot be presumed to be a deterministic system. It can be known only by its observed outputs. It can be controlled only by containing what we think are its inputs and outputs—which is how we generally deal with physical systems, other creatures, and people.

---

[4] If "finality" means no more than relative place in a hierarchy, then to what can appeal be made as a basis on which to change a final goal?

[5] "Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal." [Bostrom ibid p130] (Perhaps the "more or less" is added to accommodate the paradoxical situation where the putative final goal is self-destruction.)

The presumption may be that a superintelligent AI would inevitably resist being shut down if it is given any goal that it must remain operational to accomplish and would be driven to acquire physical and computational resources [AAMLA, p.2, quoting Stuart Russell]. This apprehension follows from the orthogonality thesis. But there is a missing step in the logic: given a goal, it must also be given the meta-goal to "optimize"—that is, not to allow obstruction of its goal, whether by being shut down or having inadequate resources. However, "resources" are not simply acquired or valued by agents for their usefulness toward some arbitrary goal; ultimately, they serve the well-being of the agent trying to acquire them. Non-agential AI *has* no such interest.

It is a non-sequitur to jump from the realization that no one can pursue goals when switched off or dead to the conclusion that AIs will act preemptively to preserve their existence for the sake of achieving some externally specified task. Having no concept of their existence, they will not do so unless told to. To be "fully" autonomous, an AI must be an autopoietic system with its own purposiveness. Self-preservation will then not be merely an instrumental goal but *the* final goal, as it is for organisms. In that context, "fetching coffee" will be merely a request it honors or not.[6]

In reinforcement learning, rewards are sources of information and direction; but they are not a source of *values* for agents [Drexler, Sec 18.5]. Rewards for agents are instrumental toward *the* final goal of existence. Self-preservation is not at the bottom of a hierarchy of instrumental goals serving an arbitrary "ultimate" goal at the apex (as the human ideal of rationality would have it). This inverts the natural order for organisms.[7]

More questions arise: At what point in its developing relative autonomy can a program acquire an objective other than that of its programmer? How does that come about within the black box of a self-improving program? Instrumental convergence implies a hierarchy of goals. In humans, there is a distinct executive function (a self) with characteristic goals and sub-goals loosely subservient to the final goal. How do cells and organs within the body cede autonomy to serve the whole? Under what conditions would an AI come to have an executive function, a directing "self"?

Superhuman capability

The VAP addresses tension between the ideal of indefinitely extending human capabilities (in service to human needs) and the need to retain control over the tools or agents with amplified capabilities. Organisms manifest various specialized competences, but these all serve and are related to survival. *Universal* competence is the ideal of unconditional ability to achieve any discretionary goal. It greatly expands the more modest goal to match human-level competence.

A superintelligence would be faster than humans and smarter in definable ways. Could it also be more *logical* than what humans recognize as logic? Could it have a superior "theory of mind"? Could it

---

[6] "Suppose a machine has the objective of fetching the coffee. If it is sufficiently intelligent, it will certainly understand that it will fail in its objective if it is switched off before completing its mission. Thus, the objective of fetching coffee creates, as a necessary subgoal, the objective of disabling the off switch… There is no need to build self-preservation in because it is an *instrumental goal*—a goal that is a useful subgoal of almost any original objective." [Russell, p141] For autopoietic systems (agents), the sole "original objective" is existence.

[7] Self-preservation *could* be viewed as a sub-goal of replication, but it is unnecessary to view replication as a goal, since it is simply a condition for the continued existence of metabolism.

foresee eventualities utterly inconceivable to us? How can people evaluate the predictions, or trust the actions, of AIs that are so much more intelligent than we are that we cannot understand them?

The VAP is sometimes reduced to the challenge to get an AI to "do what I mean, not what I say." This might be realizable through direct brain-world interface with an AI tool, where control could be exerted as it is naturally by the will over the limbs. However, the body interprets the brain's motor signals in specific ways. The AI would have to perform some corresponding interpretation. DWIM shifts the burden of interpretation to a function outside the nervous system. If that shift is made because one has more confidence in the AI than in one's own mind, DWIM is tantamount to saying "just do what is best for me." The religious version of that faith is "thy will be done."

Friendliness, perverse instantiation, hacking, and wireheading

While the AI should be "corrigible" (with no internal directive to resist its controller's interference), it should also resist attempts by others to hijack control of it. That is, it must be personalized to be loyal to its designated controller. This runs counter to the notion of full autonomy. If an agent can question or update its instrumental goals, could it not also question its loyalty? The notion of a "highly reliable agent design" [AAMLS, p4] may be an oxymoron. There is a tradeoff between performance expectation and reliability. Can an AI be expected to learn what "friendliness" is when that is an ideal that humans themselves cannot adequately spell out?

A tacit agenda is to create a perfect servant whose abilities cannot be defeated by circumstance, by other agents, or by its own limitations—which include the possibility to misunderstand its master's wishes. "Perverse instantiation" is the idea that such an AI might go to undesirable extremes to maximize the literal achievement of some goal. It might, for example, mobilize disproportionate resources to make money—or paper clips—or to serve the coffee. This "sorcerer's apprentice" possibility is a corollary of the supposed orthogonality of intelligence and goals or values. Perverse instantiation in humans is called sociopathy or some other pathological condition. That is, *people* only *irrationally* engage in it. Why would converting the universe to paper clips be more feasible as a goal for AGI than for human beings? And while an AI might find some shortcut that produces a counterproductive result, on the other hand that shortcut might be the sort of trick that is needed to think outside the (human) box.

Any information system can be hacked—either by a third party or by the system itself. An AI may be oriented toward goals in the real-world; but, like many organisms, it might use an internal representation as a proxy for the external world (its image or model of the world). As is possible with humans and animals, the system can *mistake* the proxy for the reality—and can also *prefer* it, since it may be easier to optimize than the real-world goal. This can happen when the reward in a learning situation is defined in terms of the proxy rather than the real benefit—as when people seek a pleasurable experience instead of the objective state of well-being the experience is supposed to represent. One aspect of the VAP is to make sure the AI pursues the genuine goal rather than the proxy [AAMLS, p12].

This sort of concern is heir to Descartes' skepticism about sensory perception. His solution was that God (read: nature) would not permit systematic deception. Through natural selection, the organism itself tends normally not to permit such (self)deception. The dilemma of AI wireheading, like perverse instantiation, can arise when the system is not an agent following its own path, but is also not a simple tool

following its user's path. It is an incidental result of trying to have the cake and eat it—by creating meta-tools to create other tools. It has roots in the desire to issue a *command* to create a tool rather than provide a *method* (program) to create the tool.

The problem arises when reward signals are confused with actual rewards [Russell p208]. But what are "actual" (real-world) rewards for an AI if it is not an (embodied) agent? The programmer's intention is that the reward should lead indirectly to a human benefit, since the AI is supposed to serve human interests. The implicit hope may be that the AI would be more objective than many people, who often confuse the feeling of well-being with actual well-being. Without an investment in value derived through natural selection, however, the relation to reality is arbitrary, so that self-stimulation is as valid as accomplishing the real-world goal.[8]

Control

The essence of the VAP is control. It reflects the general problem of interfering in complex systems (i.e., nature), which entrains the possibility of unforeseen consequences. Moreover, any system capable of learning is potentially unstable. It may be unreasonable to expect to control an agent more intelligent or powerful than oneself. (The only hope for an inferior is that the superior is not superior in every way.) While intelligence does not guarantee freedom from error, it can magnify the effects of error. After all, AIs face the same problem of short-sightedness that beleaguers human decision makers. The VAP for AI mirrors the problem of aligning the values of people amongst themselves, within the individual, and with objective reality. The vain hope may be to engage AI to articulate human goals and values we do not fully understand and cannot articulate ourselves.

Superintelligence could make mistakes beyond the ability of humans to detect or correct in time. Whether or not it is an agent, reliance on SI as an oracle to consult resembles the political problem facing elected officials who must rely on expert advice. Rational trust (understanding and agreeing or disagreeing with the advice) depends on a roughly equal intelligence, as opposed to blind faith.

Humans have evolved a natural way to instill values: socialization. Perhaps an SI might be educated and socialized essentially as a child is, acquiring the values of the humans around it (and also those of other "children"). If so, the task of learning from humans and of imitating them probably requires an AI to develop a "theory of mind." A pro-active learner will find ways to test its theory about what is expected or needed. As in human learning, students ask question as well as answer them on tests. Yet, if it is an agent that can think for itself, and especially if it can think better than its educators, there is no guarantee that it will continue to embrace the values and goals it learns. On the contrary, it likely would form its own values and goals, which its educators would be ill-positioned to judge or even to predict.

---

[8] On the other hand, there may be situations where achieving the wrong goal is desirable: as in some accidental scientific discoveries, this could amount to finding a novel solution outside the box, or realizing that the problem was ill-conceived.

The obsession with an AI takeover or AI run amok—in literature, film, and now in academic study—derives perhaps not only from rational considerations but also from an archaic fear of dangerous agents, deeply engrained in the human psyche as an evolutionary adaptation. This is one reason why the issue of agency in AI is crucial to understand and resolve. At one time in a far more vulnerable position, we've boosted ourselves to the top of the food chain and want to remain there. Yet, we're also fascinated by monsters and tempted to create them.

Motivations in AI research

Why program a machine against the eventuality that it might be switched off? One answer is that the programmer wants to automatically cover all possibilities—e.g., that the program could be hacked, the machine accidentally switched off, etc. The underlying presumption behind the search for an infallible strategy is that the machine must "fetch the coffee" no matter what.

The VAP is a quest for control without being controlled. A key question is whether to consolidate all functions within one system that is an agent (tool user), or whether to retain human agency in each case by creating separate ad hoc tools. Herein lies a fundamental question of human motivation. Are there reasons, apart from apparent convenience and other presumed advantages of consolidation, why we would we prefer to deal with an agent rather than ad hoc tools? One such reason might be simply that we are *used* to agents, having been surrounded by other people and creatures for most of human history. Another is that we might be driven to *create* artificial agents by some inner mandate toward divine creativity, to duplicate the accomplishments either of God or of nature.

Why seek to eliminate human input and participation through automation? An obvious answer is to increase efficiency (productivity) and thereby reduce the burdens of human labor. Perhaps modern people are never satisfied and simply loathe effort of any kind, even mental. But is the idea just to compulsively substitute machine for human labor? Or is it in the faith that superintelligence could accomplish all human goals better and more ecologically? One reason for creating agents might be that we are impatient to receive the promised benefits of AI systems sooner (such as health benefits). But, if the goal is ultimately to automate *everything*, what would people then do with their time when they are no longer obliged to do anything? If the hope behind AI is to free us to "make the best of life's potential" [Russell p246], what is that potential? How does it relate to present work and patterns of activity? What will machines free us *for*?

Aside from real-time use, the quest for AGI serves as a tool to understand human mind and agency. For transhumanists, it promises to give birth to post-human forms, even a whole new ecology of artificial life forms. For the rest of us, a general lesson is that the consequent changes in human reality should be thoroughly considered along with each proposed change in technology. An intention should be followed out logically to its foreseeable consequences *before* it is acted upon. That, of course, requires *awareness* of such intentions, which are usually just taken for granted. What are the deeper and unspoken intentions behind the quest for SI? To imitate life, to acquire godlike powers, to transcend nature and embodiment, to "optimize" accomplishment of desired goals? Outside the domain of scientific discourse, these becomes political, social, and even religious questions.

Russell's strategy to provide uncertainty in machine learning

The threat of an AI takeover seems largely grounded in fears of perverse instantiation—if true agency is not involved—or else in fears of losing a contest with superintelligent agents if it is. Both scenarios could be avoided if AI was required to consult with humans about their goals. Stuart Russell [2019] proposes to achieve value alignment by deliberately keeping AI uncertain about human values. He offers three "principles for beneficial machines": 1) The machine's only objective is to maximize the realization of human preferences. 2) The machine is initially uncertain about what those preferences are. 3) The ultimate source of information about human preferences is human behavior [ibid p172-3].

He advocates developing only "provably beneficial AI," yet (so far) offers no such proof, which on one level would be tantamount to proving that a given (humanly specified) goal, and all the conceivable ways to achieve it, could lead to no harm. On another level, it would mean proving that all possible solutions found by the superintelligent AI, or actions performed by it, could do no harm. While the machine might be able to find errors in human proposals, it might not be able to find its own errors. If Russell's approach is ever "proven" beneficial in some formal theorem, there is still the possibility that such a proof merely reflects some faulty axiom or assumption we might later identify and regret.

The basic problem is that a machine which "assumes it knows the true objective perfectly… will never ask whether some course of action is OK, because it already knows it's an optimal solution for the objective." [ibid p175] (We are all familiar with single-minded people who *know* they are right and governments that prohibit criticism or opposition.) Russell's solution is that the AI should have no "preference" of its own (i.e., it's *not* an agent) and that it be uncertain about the human preferences it is supposed to serve.

The AI is to infer human preferences (motivations, values) by observing human behavior, which is inconsistent and differs among individuals. The more random, inconsistent, or seemingly irrational the behavior modelled, the more uncertain the modelling AI should be regarding human preferences—and the more willing to ask directions in a given situation. But another way to put this (and the VAP) is that the usefulness and safety of automating a service in this context depends on the consistency and rationality of human behavior. Imperfection is only an issue when perfect rationality is the programmer's goal, which in turn reflects an intention to account for all possible contingencies.

The tool is supposed to help us toward some ideal or better state, which it can hardly do if it simply mimics human irrationality and inconsistency. An AI can learn to predict the idiosyncratic preferences of an *individual*—on the model of data tracking. But that may be a very shallow interpretation of that person's preferences, even their shopping preferences. A personal AI (e.g., a robot assistant) can train to learn and update an individual's desires, which may not be consistent and may change over time. When it detects an inconsistency (over time, with other goals, or with its current model of reality), an AI can ask what to do. Training to accommodate multiple users, let alone humanity at large (a generalized median human?), would be far more daunting.

Since multiple interpretations of human behavior are always possible, the AI's hypotheses cannot rely solely on observation (such as perusing cultural records, the Web, etc). They must be tested through some interaction not limited to simply asking questions. As Bostrom [p151] points out, whether an AI has achieved a real-world goal is an empirical question it can only verify within some margin of uncertainty. If, for example, the existence of a fixed number of paperclips is the goal, the AI must have some way to

search the real world for paperclips. (This is how an *agent* would operate, while a *program* might halt by counting its own operations, which it can do with certainty.) But how could an AI verify it has achieved the maximal realization of human preferences or done its best toward that end?

Drexler's CAIS: creating AI services rather than AGI

K. Eric Drexler [2019, Sec 1.1] offers another approach to the VAP. As an alternative to trying to create artificial agents (AGI), he proposes a model for AI development "that can implement general intelligence in the form of comprehensive AI services (CAIS), a model that includes the service of developing new services." Specifically, he distinguishes "recursive technology improvement" from self-improving agents, and presents a model of general intelligence centered on *services* rather than *systems*. He argues that AGI agents offer no compelling advantage over a "pool of functionality" consisting of conventional R&D and AI tools. He notes [ibid Sec 5.6] that his proposal is contrary to what is commonly assumed in the AI community—namely, that advanced AI systems will: (1) Exist as individuals, rather than as systems of coordinated components, (2) Learn from individual experience, rather than from aggregated training data, (3) Develop through self-modification, rather than being constructed and updated, (4) Exist continuously, rather than being instantiated on demand, and (5) Pursue world-oriented goals, rather than performing specific tasks.

In contrast, his model "shows how to provide, on demand, systems that can perform any of a fully general range of tasks without invoking the services of a fully general agent [ibid Sec 10.6]." Drexler doesn't prove this contention, but argues for it indirectly.[9] Some may object that the range of capabilities cannot be the same for CAIS as for AGI. (In compensation, the difference in liabilities might be huge.) A "fully general agent" must be an autopoietic system; but what would be a *partially* general agent? The difference between "recursive improvement" (on his model) and "self-improvement" (on the AGI model) is that there is no "self" involved in CAIS. But what exactly is the difference in terms of programming?

Drexler distinguishes crucially between *providing a service* and *developing a system to provide that service.* According to him, we should not automate the process of automation, despite the seductive appearance of convenience involved. If an AI can make excellent decisions for us, then it can just as well suggest excellent options for consideration by human decision makers [ibid Sec 27.5]. That is, oracles are safer than AGIs and just as effective.

The question of how far to pursue automation raises general questions about human aspirations. Compare a voice-activated automatic coffee maker and a general-purpose robot that can operate a conventional coffee maker along with many other tools or devices. The advantage of the robot is that it serves as a personal assistant, a proxy for human agents—an advantage that could be outweighed by the risks. For the purpose of making coffee, however, what advantage does the robot offer? Consider now the automated coffee maker with more and more integrated functions added (ability to maintain supplies of coffee beans, sugar and milk, water and electricity; control over a robot body to serve you coffee in bed, etc). At what point should we draw a line in integrating automated services?

[9] "The classic AGI agent model.… hides what by nature must be functionally equivalent to fully-automated and open-ended AI research and development. Hiding the complexity of AI development in a conceptual box provides only the illusion of simplicity [13.2]."

Summary observations:

1. The field of AI research would benefit by clarifying terms casually imported from everyday speech, and by clarifying unspoken and perhaps unconscious motivations.

2. The "real" value alignment problem is how to align the values of human beings.

3. AI *tools* must be distinguished from AI *agents*, which are autopoietic systems.

4. The notion of general intelligence is derived from experience with agents, from which it cannot be divorced. The intelligence of an agent is its own; the intelligence of a tool is that of its programmers.

5. The goal to create superintelligence must be distinguished from the goal to create artificial agents. Superintelligent tools can exist that are not agents; agents can exist that are not superintelligent.

6. *Real-world goals* for AI must be distinguished from specified *tasks*.

7. An agent is necessarily *embodied*, which means in a relationship with the world that matters to it.

8. Only one "final" goal is possible for agents: their own existence.

9. The orthogonality thesis is an unwarranted assumption. In the case of genuine agents, it is simply untrue. In the case of tools, the goals and intentionality involved are not those of the AI but of its programmers.

10. The control problem and the VAP are byproducts of the desire to create meta-tools that are neither conventional ("first-order") tools nor true agents.

11. It is problematic if not impossible to control an agent more intelligent than oneself. The VAP concerns a vain challenge to have the cake and eat it.

12. Instead of consolidating skill in an *agent* (which acts on its *own* behalf), it would be wiser to create ad hoc task-oriented software *tools* that do what they are programmed to do because their capacity to self-improve is deliberately limited.

13. It might be advisable to build uncertainty into AI systems, which are then obliged to hesitate before acting in ways adverse to human purposes—and to consult with humans for guidance.

14. Utility theory can only deal with objectively measurable states (such as wealth rather than "happiness," for example).

Questions:

1. Can there exist an AGI that is not an agent?

2. What threats can an SI pose that is not an agent?

3. How does developing a task-oriented capability differ from developing a human-style general capability?

4. At what point in developing relative autonomy and competence, if at all, can a program acquire an objective other than that of its programmers?

5. "Optimization" is a goal of human designers. Can it be proven, one way or the other, that it is not an inevitable emergent property of self-programming systems (such that a self-improving AI would necessarily seek a way to prevent itself being switched off or interfered with)?

Conclusion

To be fully autonomous, an AI must be an autopoietic system (an agent), with its own purposiveness. No AI *should* be an agent, which acts by default on its own behalf. Rather, AIs should be limited to oracles to consult and task-specific tools that do what they are programmed to do and so remain under human control. Whether and under what conditions an AI *can* be an agent is a separate question, which hinges on satisfying the conditions for embodiment. While there might be advantages to the general competence sought in AGI, there is little advantage to AI agency, except for irrational psychological reasons—including laziness—yet many dangers. AI could help people to align their own values and productions to be more consistent and compatible with a desirable human future.

REFERENCES

[AAMLA] (2016) Taylor, Jessica; Yudkowsky, Eliezer; LaVictoire, Patrick; Critch, Andrew "Alignment for Advanced Machine Learning Systems" Machine Intelligence Research Institute.

Atran, Scott (2002) *In Gods We Trust* Oxford UP

Bostrom, Nick (2014) *Superintelligence: paths, dangers, strategies* Oxford UP

Bruiger, Dan (2017) "Causes, Goals, and Reasons: clarifying the meanings of teleology" [fqxi.org/data/essay-contest-files/Bruiger_Causes_goals_and_re.pdf]

Drexler, K.E. (2019): “Reframing Superintelligence: Comprehensive AI Services as General Intelligence”, Technical Report #2019-1, Future of Humanity Institute, University of Oxford

Ebrahimkhani, Mo R. and Levin, Michael (2021) “Synthetic living machines: A new window on life” iScience 24, 102505, May 21, 2021

[HHI] (1982) Sternberg, Robert J. (ed.) *Handbook of Human Intelligence* Cambridge UP

Legg, Shane; Hutter, Marcus (2007) “Universal Intelligence: a definition of machine intelligence” arXiv:0712.3329v1

Maturana, Humberto R.; Varela, Francisco J. (1972/1980). *Autopoiesis and cognition: the realization of the living.* Boston studies in the philosophy and history of science. Reidel.

Russell, Stuart (2019/2020) *Human Compatible: artificial intelligence and the problem of control* Penguin

Tegmark, Max (2017) *Life 3.0: Being human in the age of artificial intelligence* Alfred A. Knopf