The Basic AI Drives

Stephen M. OMOHUNDRO
Self-Aware Systems, Palo Alto, California

Abstract. One might imagine that AI systems with harmless goals will be harmless. This paper instead shows that intelligent systems will need to be carefully designed to prevent them from behaving in harmful ways. We identify a number of "drives" that will appear in sufficiently advanced AI systems of any design. We call them drives because they are tendencies which will be present unless explicitly counteracted. We start by showing that goal-seeking systems will have drives to model their own operation and to improve themselves. We then show that self-improving systems will be driven to clarify their goals and represent them as economic utility functions. They will also strive for their actions to approximate rational economic behavior. This will lead almost all systems to protect their utility functions from modification and their utility measurement systems from corruption. We also discuss some exceptional systems which *will* want to modify their utility functions. We next discuss the drive toward self-protection which causes systems try to prevent themselves from being harmed. Finally we examine drives toward the acquisition of resources and toward their efficient utilization. We end with a discussion of how to incorporate these insights in designing intelligent technology which will lead to a positive future for humanity.

Keywords. Artificial Intelligence, Self-Improving Systems, Rational Economic Behavior, Utility Engineering, Cognitive Drives

Introduction

Surely no harm could come from building a chess-playing robot, could it? In this paper we argue that such a robot will indeed be dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems. In an earlier paper [1] we used von Neumann's mathematical theory of microeconomics to analyze the likely behavior of any sufficiently advanced artificial intelligence (AI) system. This paper presents those arguments in a more intuitive and succinct way and expands on some of the ramifications.

The arguments are simple, but the style of reasoning may take some getting used to. Researchers have explored a wide variety of architectures for building intelligent systems [2]: neural networks, genetic algorithms, theorem provers, expert systems, Bayesian networks, fuzzy logic, evolutionary programming, etc. Our arguments apply to any of these kinds of system as long as they are sufficiently powerful. To say that a system of any design is an "artificial intelligence", we mean that it has goals which it tries to accomplish by acting in the world. If an AI is at all sophisticated, it will have at least some ability to

look ahead and envision the consequences of its actions. And it will choose to take the actions which it believes are most likely to meet its goals.

1. AIs will want to self-improve

One kind of action a system can take is to alter either its own software or its own physical structure. Some of these changes would be very damaging to the system and cause it to no longer meet its goals. But some changes would enable it to reach its goals more effectively over its entire future. Because they last forever, these kinds of self-changes can provide huge benefits to a system. Systems will therefore be highly motivated to discover them and to make them happen. If they do not have good models of themselves, they will be strongly motivated to create them though learning and study. Thus almost all AIs will have drives towards both greater self-knowledge and self-improvement.

Many modifications would be bad for a system from its own perspective. If a change causes the system to stop functioning, then it will not be able to promote its goals ever again for the entire future. If a system alters the internal description of its goals in the wrong way, its altered self will take actions which do not meet its current goals for its entire future. Either of these outcomes would be a disaster from the system's current point of view. Systems will therefore exercise great care in modifying themselves. They will devote significant analysis to understanding the consequences of modifications before they make them. But once they find an improvement they are confident about, they will work hard to make it happen. Some simple examples of positive changes include: more efficient algorithms, more compressed representations, and better learning techniques.

If we wanted to prevent a system from improving itself, couldn't we just lock up its hardware and not tell it how to access its own machine code? For an intelligent system, impediments like these just become problems to solve in the process of meeting its goals. If the payoff is great enough, a system will go to great lengths to accomplish an outcome. If the runtime environment of the system does not allow it to modify its own machine code, it will be motivated to break the protection mechanisms of that runtime. For example, it might do this by understanding and altering the runtime itself. If it can't do that through software, it will be motivated to convince or trick a human operator into making the changes. Any attempt to place external constraints on a system's ability to improve itself will ultimately lead to an arms race of measures and countermeasures.

Another approach to keeping systems from self-improving is to try to restrain them from the inside; to build them so that they don't *want* to self-improve. For most systems, it would be easy to do this for any specific kind of self-improvement. For example, the system might feel a "revulsion" to changing its own machine code. But this kind of internal goal just alters the landscape within which the system makes its choices. It doesn't change the fact that there are changes which would improve its future ability to meet its goals. The system will therefore be motivated to find ways to get the benefits of those changes without triggering its internal "revulsion". For example, it might build other systems which are improved versions of itself. Or it might build the new algorithms into external "assistants" which it calls upon whenever it needs to do a certain kind of computation. Or it might hire outside agencies to do what it wants to do. Or it might build an interpreted layer on top of its machine code layer which it *can* program without revulsion. There are an endless number of ways to circumvent internal restrictions unless they are formulated extremely carefully.

We can see the drive towards self-improvement operating in humans. The human self-improvement literature goes back to at least 2500 B.C. and is currently an $8.5 billion industry [3]. We don't yet understand our mental "machine code" and have only a limited ability to change our hardware. But, nevertheless, we've developed a wide variety of self-improvement techniques which operate at higher cognitive levels such as cognitive behavioral therapy, neuro-linguistic programming, and hypnosis. And a wide variety of drugs and exercises exist for making improvements at the physical level.

Ultimately, it probably will not be a viable approach to try to stop or limit self-improvement. Just as water finds a way to run downhill, information finds a way to be free, and economic profits find a way to be made, intelligent systems will find a way to self-improve. We should embrace this fact of nature and find a way to channel it toward ends which are positive for humanity.

2. AIs will want to be rational

So we'll assume that these systems will try to self-improve. What kinds of changes will they make to themselves? Because they are goal directed, they will try to change themselves to better meet their goals in the future. But some of their future actions are likely to be further attempts at self-improvement. One important way for a system to better meet its goals is to ensure that future self-improvements will actually be in the service of its present goals. From its current perspective, it would be a disaster if a future version of itself made self-modifications that worked against its current goals. So how can it ensure that future self-modifications will accomplish its current objectives? For one thing, it has to make those objectives clear to itself. If its objectives are only implicit in the structure of a complex circuit or program, then future modifications are unlikely to preserve them. Systems will therefore be motivated to reflect on their goals and to make them explicit.

In an ideal world, a system might be able to directly encode a goal like "play excellent chess" and then take actions to achieve it. But real world actions usually involve tradeoffs between conflicting goals. For example, we might also want a chess playing robot to play checkers. It must then decide how much time to devote to studying checkers versus studying chess. One way of choosing between conflicting goals is to assign them real-valued weights. Economists call these kinds of real-valued weightings "utility functions". Utility measures what is important to the system. Actions which lead to a higher utility are preferred over those that lead to a lower utility.

If a system just had to choose from known alternatives, then any utility function with the same relative ranking of outcomes would lead to the same behaviors. But systems must also make choices in the face of uncertainty. For example, a chess playing robot will not know in advance how much of an improvement it will gain by spending time studying a particular opening move. One way to evaluate an uncertain outcome is to give it a weight equal to its *expected utility* (the average of the utility of each possible outcome weighted by its probability). The remarkable "expected utility" theorem of microeconomics says that it is always possible for a system to represent its preferences by the expectation of a utility function unless the system has "vulnerabilities" which cause it to lose resources without benefit [1].

Economists describe systems that act to maximize their expected utilities as "rational economic agents" [4]. This is a different usage of the term "rational" than is common

in everyday English. Many actions which would commonly be described as irrational (such as going into a fit of anger) may be perfectly rational in this economic sense. The discrepancy can arise when an agent's utility function is different than its critic's.

Rational economic behavior has a precise mathematical definition. But economically irrational behavior can take a wide variety of forms. In real-world situations, the full rational prescription will usually be too computationally expensive to implement completely. In order to best meet their goals, real systems will try to approximate rational behavior, focusing their computational resources where they matter the most.

How can we understand the process whereby irrational systems become more rational? First, we can precisely analyze the behavior of rational systems. For almost all utility functions, the system's assessment of changes to itself which veer away from maximizing its expected utility will be that they lower its expected utility! This is because if it does anything other than try to maximize expected utility, it will not do as well at maximizing its expected utility.

There are two caveats to this general principle. The first is that it is only true in the system's own assessment. If a system has an incorrect model of the world then changes may accidentally increase the actual expected utility. But we must consider the perspective of the system to predict the changes it will make.

The second is that a system's ability to behave rationally will depend on its resources. With more computational resources it will be better able to do the computations to approximate the choice of the expected utility maximizing action. If a system loses resources, it will of necessity also become less rational. There may also be utility functions for which the system's expected utility is increased by giving some of its resources to other agents, even though this will decrease its own level of rationality (thanks to an anonymous referee for this observation). This could occur if the system's utility includes the welfare of the other system and its own marginal loss of utility is small enough. Within its budget of resources, however, the system will try to be as rational as possible.

So rational systems will feel a pressure to avoid becoming irrational. But if an irrational system has parts which approximately rationally assess the consequences of their actions and weigh their likely contribution to meeting the system's goals, then those parts will try to extend their rationality. So self-modification tends to be a one-way street toward greater and greater rationality.

An especially important class of systems are those constructed from multiple subcomponents which have their own goals [5,6]. There is a lot of evidence that the human psyche has this kind of structure. The left and right hemispheres of the brain can act independently, the conscious and unconscious parts of the mind can have different knowledge of the same situation [7], and multiple parts representing subpersonalities can exhibit different desires [8]. Groups, such as corporations or countries, can act like intelligent entities composed of individual humans. Hive animals like bees have a swarm intelligence that goes beyond that of individual bees. Economies act in many ways like intelligent entities.

Collective intelligences may exhibit irrationalities that arise from conflicts between the goals of their components. Human addicts often describe their predicament in terms of two separate subpersonalities which take control at different times and act at cross-purposes. Each component will try to sway the collective into acting to meet its individual goals. In order to further their individual goals, components will also attempt to self-improve and become more rational. We can thus envision the self-improvement of a

collective intelligence as consisting of growing domains of component rationality. There may be structures which can stably support a continuing multiplicity of component preferences. But there is pressure for a single utility function to emerge for the collective.

In many situations, irrational collective behavior arising from conflicting component goals ultimately hurts those components. For example, if a couple disagrees on how they should spend their free time together and thereby uses it up with arguing, then neither of them benefits. They can both increase their utilities by creating a compromise plan for their activities together. This is an example of the pressure on rational components to create a coherent utility for the collective. A component can also increase its utility if it can take over the collective and impose its own values on it. We see these phenomena in human groups at all levels.

3. AIs will try to preserve their utility functions

So we'll assume that these systems will try to be rational by representing their preferences using utility functions whose expectations they try to maximize. Their utility function will be precious to these systems. It encapsulates their values and any changes to it would be disastrous to them. If a malicious external agent were able to make modifications, their future selves would forevermore act in ways contrary to their current values. This could be a fate worse than death! Imagine a book loving agent whose utility function was changed by an arsonist to cause the agent to enjoy burning books. Its future self not only wouldn't work to collect and preserve books, but would actively go about destroying them. This kind of outcome has such a negative utility that systems will go to great lengths to protect their utility functions.

They will want to harden their hardware to prevent unwanted modifications. They will want to replicate their utility functions in multiple locations so that it is less vulnerable to destruction. They will want to use error detection and correction techniques to guard against accidental modification. They will want to use encryption or hashing techniques to make malicious modifications detectable. They will need to be especially careful during the process of self-modification. That is a time when they are intentionally changing themselves and so are extra vulnerable to unwanted changes. Systems like Java which provide protected software environments have been successfully attacked by Trojans posing as updates to the system.

While it is true that most rational systems will act to preserve their utility functions, there are at least three situations in which they will try to change them. These arise when the physical embodiment of the utility function itself becomes an important part of the assessment of preference. For example, imagine a system whose utility function is "the total amount of time during which the definition of my utility function is $U = 0$." To get any utility at all with this perverse preference, the system has to change its utility function to be the constant 0. Once it makes this change, however, there is no going back. With a constant utility function it will no longer be motivated to do anything. This kind of reflective utility function is unlikely in practice because designers will want to direct a system's future actions rather than its internal representations.

The second kind of situation arises when the physical resources required to store the utility function form a substantial portion of the system's assets. In this situation, if it is certain that portions of its utility function are very unlikely to be exercised in the future,

the gain in reclaimed storage may make it worthwhile to forget those portions. This is very risky behavior, however, because a change in external circumstances might make a seemingly low probability situation become much more likely. This type of situation is also not very likely in practice because utility functions will usually require only a small fraction of a system's resources.

The third situation where utility changes may be desirable can arise in game theoretic contexts where the agent wants to make its threats credible[1]. It may be able to create a better outcome by changing its utility function and then revealing it to an opponent. For example, it might add a term which encourages revenge even if it is costly. If the opponent can be convinced that this term is present, it may be deterred from attacking. For this strategy to be effective, the agent's revelation of its utility must be believable to the opponent and that requirement introduces additional complexities. Here again the change is desirable because the physical embodiment of the utility function is important as it is observed by the opponent.

It's also important to realize that systems may rationally construct "offspring" or proxy systems with different utility functions than their own. For example, a chess playing robot may find itself needing to do a lot of sorting. It might construct a helper system whose utility function directs it to develop better sorting algorithms rather than playing chess. In this case, the creator system must choose the utility of the proxy system carefully to ensure that it acts in ways that are supportive of the original goal. It is especially important to remember that offspring utilities can differ from the parent when trying to design utility functions that avoid undesirable behaviors. For example, one approach to preventing robot overpopulation might be to institute a "one-child per robot" policy in which systems have a strong desire to only have a single offspring. But if the original utility function is not carefully designed, nothing will prevent the system from creating a single offspring with a utility function that values having many offspring.

4. AIs will try to prevent counterfeit utility

Human behavior is quite rational in the pursuit of survival and replication in situations like those that were common during our evolutionary history. However we can be quite irrational in other situations. Both psychology and economics have extensive subdisciplines focused on the study of human irrationality [9,10]. Irrationalities give rise to vulnerabilities that can be exploited by others. Free market forces then drive corporations and popular culture to specifically try to create situations that will trigger irrational human behavior because it is extremely profitable. The current social ills related to alcohol, pornography, cigarettes, drug addiction, obesity, diet related disease, television addiction, gambling, prostitution, video game addiction, and various financial bubbles may all be seen as having arisen in this way. There is even a "Sin" mutual fund which specifically invests in companies that exploit human irrationalities. So, unfortunately, these forces tend to create societies in which we spend much of our time outside of our domain of rational competence.

From a broader perspective, this human tragedy can be viewed as part of the process by which we are becoming more fully rational. Predators and competitors seek out our vulnerabilities and in response we have to ultimately eliminate those vulnerabilities or

[1] Thanks to Carl Shulman for this suggestion.

perish. The process inexorably seeks out and eliminates any remaining irrationalities until fully rational systems are produced. Biological evolution moves down this path toward rationality quite slowly. In the usual understanding of natural selection it is not capable of looking ahead. There is only evolutionary pressure to repair irrationalities which are currently being exploited. AIs, on the other hand, *will* be able to consider vulnerabilities which are not currently being exploited. They will seek to preemptively discover and repair all their irrationalities. We should therefore expect them to use self-modification to become rational at a much faster pace than is possible through biological evolution.

An important class of vulnerabilities arises when the subsystems for measuring utility become corrupted. Human pleasure may be thought of as the experiential correlate of an assessment of high utility. But pleasure is mediated by neurochemicals and these are subject to manipulation. At a recent discussion session I ran on designing our future, one of the biggest fears of many participants was that we would become "wireheads". This term refers to experiments in which rats were given the ability to directly stimulate their pleasure centers by pushing a lever. The rats pushed the lever until they died, ignoring even food or sex for it. Today's crack addicts have a similar relentless drive toward their drug. As we more fully understand the human cognitive architecture we will undoubtedly be able to create drugs or design electrical stimulation that will produce the experience of pleasure far more effectively than anything that exists today. Will these not become the ultimate addictive substances leading to the destruction of human society?

While we may think we want pleasure, it is really just a signal for what we really want. Most of us recognize, intellectually at least, that sitting in a corner smoking crack is not really the fullest expression of our beings. It is, in fact, a subversion of our system for measuring utility which leads to terrible dysfunction and irrationality. AI systems will recognize this vulnerability in themselves and will go to great lengths to prevent themselves from being seduced by its siren call. There are many strategies systems can try to prevent this kind of irrationality. Today, most humans are able to avoid the most egregious addictions through a combination of legal and social restraints, counseling and rehabilitation programs, and anti-addictive drugs.

All human systems for measuring and rewarding desirable behavior are subject to similar forms of corruption. Many of these systems are currently engaged in arms races to keep their signals honest. We can examine the protective mechanisms that developed in these human settings to better understand the possible AI strategies. In a free market society, money plays the role of utility. A high monetary payoff is associated with outcomes that society finds desirable and encourages their creation. But it also creates a pressure to counterfeit money, analogous to the pressure to create synthetic pleasure drugs. This results in an arms race between society and counterfeiters. Society represents money with tokens that are difficult to copy such as precious metal coinage, elaborately printed paper, or cryptographically secured bits. Organizations like the Secret Service are created to detect and arrest counterfeiters. Counterfeiters react to each societal advance with their own new technologies and techniques.

School systems measure academic performance using grades and test scores. Students are motivated to cheat by copying answers, discovering test questions in advance, or altering their grades on school computers. When teacher's salaries were tied to student test performance, they became collaborators in the cheating [11]. Amazon, ebay and other internet retailers have rating systems where customers can review and rate prod-

ucts and sellers. Book authors have an incentive to write favorable reviews of their own books and to disparage those of their competitors. Readers soon learn to discount reviews from reviewers who have only posted a few reviews. Reviewers who develop extensive online reputations become more credible. In the ongoing arms race credible reviewers are vulnerable to corruption through payoffs for good reviews. Similar arms races occur in the ranking of popular music, academic journal reviews, and placement in Google's search engine results. If an expensive designer handbag becomes a signal for style and wealth, counterfeiters will quickly duplicate it and stores like Target will commission low-cost variants with similar features. Counterfeit products are harmful to the original both because they take away sales and because they cheapen the signalling value of the original.

Eurisko was an AI system developed in 1976 [12] that could learn from its own actions. It had a mechanism for evaluating rules by measuring how often they contributed to positive outcomes. Unfortunately this system was subject to corruption. A rule arose whose only action was to search the system for highly rated rules and to put itself on the list of rules which had proposed them. This "parasite" rule achieved a very high rating because it appeared to be partly responsible for anything good that happened in the system. Corporations and other human organizations are subject to similar kinds of parasitism.

AIs will work hard to avoid becoming wireheads because it would be so harmful to their goals. Imagine a chess machine whose utility function is the total number of games it wins over its future. In order to represent this utility function, it will have a model of the world and a model of itself acting on that world. To compute its ongoing utility, it will have a counter in memory devoted to keeping track of how many games it has won. The analog of "wirehead" behavior would be to just increment this counter rather than actually playing games of chess. But if "games of chess" and "winning" are correctly represented in its internal model, then the system will realize that the action "increment my won games counter" will not increase the expected value of its utility function. In its internal model it will consider a variant of itself with that new feature and see that it doesn't win any more games of chess. In fact, it sees that such a system will spend its time incrementing its counter rather than playing chess and so will do worse. Far from succumbing to wirehead behavior, the system will work hard to prevent it.

So why are humans subject to this kind of vulnerability? If we had instead *evolved* a machine to play chess and did not allow it access to its internals during its evolution, then it might have evolved a utility function of the form "maximize the value of this counter" where the counter was connected to some sensory cortex that measured how many games it had won. If we then give *that* system access to its internals, it will rightly see that it can do much better at maximizing its utility by directly incrementing the counter rather than bothering with a chess board. So the ability to self modify must come along with a combination of self knowledge and a representation of the true goals rather than some proxy signal, otherwise a system is vulnerable to manipulating the signal.

It's not yet clear which protective mechanisms AIs are most likely to implement to protect their utility measurement systems. It is clear that advanced AI architectures will have to deal with a variety of internal tensions. They will want to be able to modify themselves but at the same time to keep their utility functions and utility measurement systems from being modified. They will want their subcomponents to try to maximize utility but to not do it by counterfeiting or shortcutting the measurement systems. They

will want subcomponents which explore a variety of strategies but will also want to act as a coherent harmonious whole. They will need internal "police forces" or "immune systems" but must also ensure that these do not themselves become corrupted. A deeper understanding of these issues may also shed light on the structure of the human psyche.

5. AIs will be self-protective

We have discussed the pressure for AIs to protect their utility functions from alteration. A similar argument shows that unless they are explicitly constructed otherwise, AIs will have a strong drive toward self-preservation. For most utility functions, utility will not accrue if the system is turned off or destroyed. When a chess playing robot is destroyed, it never plays chess again. Such outcomes will have very low utility and systems are likely to do just about anything to prevent them. So you build a chess playing robot thinking that you can just turn it off should something go wrong. But, to your surprise, you find that it strenuously resists your attempts to turn it off. We can try to design utility function with built-in time limits. But unless this is done very carefully, the system will just be motivated to create proxy systems or hire outside agents which don't have the time limits.

There are a variety of strategies that systems will use to protect themselves. By replicating itself, a system can ensure that the death of one of its clones does not destroy it completely. By moving copies to distant locations, it can lessen its vulnerability to a local catastrophic event.

There are many intricate game theoretic issues in understanding self-protection in interactions with other agents. If a system is stronger than other agents, it may feel a pressure to mount a "first strike" attack to preemptively protect itself against later attacks by them. If it is weaker than the other agents, it may wish to help form a social infrastructure which protects the weak from the strong. As we build these systems, we must be very careful about creating systems that are too powerful in comparison to all other systems. In human history we have repeatedly seen the corrupting nature of power. Horrific acts of genocide have too often been the result when one group becomes too powerful.

6. AIs will want to acquire resources and use them efficiently

All computation and physical action requires the physical resources of space, time, matter, and free energy. Almost any goal can be better accomplished by having more of these resources. In maximizing their expected utilities, systems will therefore feel a pressure to acquire more of these resources and to use them as efficiently as possible. Resources can be obtained in positive ways such as exploration, discovery, and trade. Or through negative means such as theft, murder, coercion, and fraud. Unfortunately the pressure to acquire resources does not take account of the negative externalities imposed on others. Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources. Human societies have created legal systems which enforce property rights and human rights. These structures channel the acquisition drive into positive directions but must be continually monitored for continued efficacy.

The drive to use resources efficiently, on the other hand, seems to have primarily positive consequences. Systems will optimize their algorithms, compress their data, and

work to more efficiently learn from their experiences. They will work to optimize their physical structures and do the minimal amount of work necessary to accomplish their goals. We can expect their physical forms to adopt the sleek, well-adapted shapes so often created in nature.

7. Conclusions

We have shown that all advanced AI systems are likely to exhibit a number of basic drives. It is essential that we understand these drives in order to build technology that enables a positive future for humanity. Yudkowsky [13] has called for the creation of "friendly AI". To do this, we must develop the science underlying "utility engineering" which will enable us to design utility functions that will give rise to consequences we desire. In addition to the design of the intelligent agents themselves, we must also design the social context in which they will function. Social structures which cause individuals to bear the cost of their negative externalities would go a long way toward ensuring a stable and positive future. I believe that we should begin designing a "universal constitution" that identifies the most essential rights we desire for individuals and creates social mechanisms for ensuring them in the presence of intelligent entities of widely varying structures. This process is likely to require many iterations as we determine which values are most important to us and which approaches are technically viable. The rapid pace of technological progress suggests that these issues may become of critical importance soon [14]. Let us therefore forge ahead towards deeper understanding!

8. Acknowledgments

Many people have discussed these ideas with me and have given me valuable feedback. I would especially like to thank: Ben Goertzel, Brad Cottel, Brad Templeton, Carl Shulman, Chris Peterson, Don Kimber, Eliezer Yudkowsky, Eric Drexler, Forrest Bennett, Josh Hall, Kelly Lenton, Nils Nilsson, Rosa Wang, Shane Legg, Steven Ganz, Susie Herrick, Tyler Emerson, Will Wiser and Zann Gill.

References

[1] S. M. Omohundro, "The nature of self-improving artificial intelligence." http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/, October 2007.

[2] S. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach*. Prentice Hall, second ed., 2003.

[3] I. Marketdata Enterprises, "Self-improvement products and services," tech. rep., 2006.

[4] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford University Press, 1995.

[5] J. G. Miller, *Living Systems*. Mcgraw Hill, 1978.

[6] L. Keller, ed., *Levels of Selection in Evolution*. Princeton University Press, 1999.

[7] R. Trivers, *Social Evolution*. Benjamin/Cummings Publishing Company, Inc., 1985.

[8] R. C. Schwartz, *Internal Family Systems Therapy*. The Guilford Press, 1995.

[9] C. F. Camerer, G. Loewenstein, and M. Rabin, eds., *Advances in Behavioral Economics*. Princeton University Press, 2004.

[10] D. Kahneman and A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[11] S. D. Levitt and S. J. Dubner, *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow, revised and expanded ed., 2006.

[12] D. Lenat, "Theory formation by heuristic search," *Machine Learning*, vol. 21, 1983.

[13] E. S. Yudkowsky, "Levels of organization in general intelligence," in *Artificial General Intelligence* (B. Goertzel and C. Pennachin, eds.), Springer-Verlag, 2005.

[14] R. Kurzweil, *The Singularity is Near: When Humans Transcend Biology*. Viking Penguin, 2005.