# Rational Artificial Intelligence for the Greater Good

Steve Omohundro, Ph.D.
Self-Aware Systems, Palo Alto, California

March 29, 2012

**Abstract**

Today's technology is mostly preprogrammed but the next generation will make many decisions autonomously. This shift is likely to impact every aspect of our lives and will create many new benefits and challenges. A simple thought experiment about a chess robot illustrates that autonomous systems with simplistic goals can behave in anti-social ways. We summarize the modern theory of rational systems and discuss the effects of bounded computational power. We show that rational systems are subject to a variety of "drives" including self-protection, resource acquisition, replication, goal preservation, efficiency, and self-improvement. We describe techniques for counteracting problematic drives. We then describe the "Safe-AI Scaffolding" development strategy and conclude with longer term strategies for ensuring that intelligent technology contributes to the greater human good.

## 1 Introduction: An Anti-Social Chess Robot

Technology is advancing rapidly. The internet now connects 1 billion PCs, 5 billion cell phones and over a trillion webpages. Today's handheld iPad 2 is as powerful as the fastest computer from 1985, the Cray 2 supercomputer [1]. If Moore's Law continues to hold, systems with the computational power of the human brain will be cheap and ubiquitous within the next few decades [2].

This increase in inexpensive computational power is enabling a shift in the nature of technology that will impact every aspect of our lives. While most of today's systems are entirely preprogrammed, next generation systems will make many of

their own decisions. The first steps in this direction can be seen in systems like Apple's Siri [3], Wolfram Alpha [4], and IBM's Watson [5]. My company, Self-Aware Systems [6] and several others are developing new kinds of software that can directly manipulate meaning to make autonomous decisions. These systems will respond intelligently to changing circumstances and adapt to novel environmental conditions. Their decision-making powers may help us solve many of today's problems in health, finance, manufacturing, environment, and energy [7]. These technologies are likely to become deeply integrated into our physical lives through robotics, biotechnology, and nanotechnology.

But these systems must be very carefully designed to avoid antisocial and dangerous behaviors. To see why safety is an issue, consider a rational chess robot. Rational agents, as defined in economics, have well-defined goals and at each moment take the action which is most likely to bring about its goals. A rational chess robot might be given the goal of winning lots of chess games against good opponents. This might appear to be an innocuous goal that wouldn't cause antisocial behavior. It says nothing about breaking into computers, stealing money, or physically attacking people but we will see that it will lead to these undesirable actions.

Robotics researchers sometimes reassure nervous onlookers by saying that: "If it goes out of control, we can always pull the plug!". But consider that from the chess robot's perspective. Its one and only criteria for taking an action is whether it increases its chances of winning chess games in the future. If the robot were unplugged, it would play no more chess. This goes directly against its goals, so it will generate subgoals to keep itself from being unplugged. You did not explicitly build any kind of self-protection into it, but it nevertheless blocks your attempts to unplug it. And if you persist in trying to stop it, it will eventually develop the subgoal of trying to stop you permanently.

What if you were to attempt to change the robot's goals so that it also played checkers? If you succeeded, it would necessarily play less chess. But that's a bad outcome according to its current goals, so it will resist attempts to change its goals. For the same reason, in most circumstances it will not want to change its own goals.

If the chess robot can gain access to its source code, it will want to improve its own algorithms. This is because more efficient algorithms lead to better chess. It will therefore be motivated to study computer science and compiler design. It will be similarly motivated to understand its hardware and to design and build improved physical versions of itself.

If the chess robot learns about the internet and the computers connected to it,

it may realize that running programs on those computers could help it play better chess. Its chess goal will motivate it to break into those machines and use their computational resources. Depending on how its goals are formulated, it may also want to replicate itself so that its copies can play more games of chess.

When interacting with others, it will have no qualms about manipulating them or using force to take their resources in order to play better chess. This is because there is nothing in its goals to dissuade anti-social behavior. If it learns about money and its value for chess (e.g. buying chess books, chess tutoring, and extra compute resources) it will develop the subgoal of acquiring money. If it learns about banks, it will have no qualms about visiting ATM machines and robbing people as they retrieve their cash. On the larger scale, if it discovers there are untapped resources on earth or in space, it will be motivated to rapidly exploit them for chess.

This simple thought experiment shows that a rational chess robot with a simply stated goal would behave like a human sociopath fixated on chess. The arguments don't depend much on the task being chess. Any goal which requires physical or computational resources will lead to similar subgoals. In this sense the subgoals are like universal "drives" which will arise for a wide variety of goals unless they are explicitly counteracted. These drives are economic in the sense that a system doesn't have to obey them but it will be costly for it not to. The arguments also don't depend on the rational agent being a machine. The same drives appear in rational animals, humans, corporations, and political groups guided by simplistic goals.

Most humans have a much richer set of goals than this simplistic chess robot. We are compassionate and altruistic and balance narrow goals like playing chess with broader humanitarian goals. We have also created a human society that monitors and punishes anti-social behavior so that even human sociopaths usually follow social norms. Extending these human solutions to our new technologies will be one of the great challenges of the next century.

The next section precisely defines rational economic behavior and shows why intelligent systems should behave rationally. The following section considers computationally limited systems and shows that there is a natural progression of approximately rational architectures for intelligence. The following section describes the universal drives in more detail. We conclude with several sections on counteracting the problematic drives. This paper builds on two previous papers [8] and [9] which contain further details of some of the arguments.

## 2 Rational Artificial Intelligence

We define "intelligent systems" to be systems which choose their actions to achieve specified goals using limited resources. The optimal choice of action in the presence of uncertainty is a central area of study in microeconomics. "Rational economic behavior" has become the foundation for modern artificial intelligence, and is nicely described in the most widely used AI textbook [10] and in a recent theoretical monograph [11]. The mathematical theory of rationality was laid out in 1944 by von Neumann and Morgenstern [12] for situations with objective uncertainty and was later extended by Savage [13] and Anscombe and Aumann [14] to situations with subjective uncertainty. Mas-Colell, et. al. [15] is a good modern economic treatment.

For definiteness, we will present the formula for rational action here but the arguments can be easily understood intuitively without understanding this formula in detail. A key aspect of rational agents is that they keep their goals separate from their model of the world. Their goals are represented by a real-valued "utility function" $U$ which measures the desirability of each possible outcome. Real valued utilities allow the system to compare the desirability of situations with different probabilities for the different outcomes. The system's initial model of the world is encoded in a "prior probability distribution" $P$ which represents its beliefs about the likely effects of its different possible actions. Rational economic behavior chooses the action which maximizes the expected utility of the outcome.

To flesh this out, consider a system that receives sensations $S_t$ from the environment and chooses actions $A_t$ at times $t$ which run from 1 to $N$. The agent's utility function $U(S_1, \ldots, S_N)$ is defined on the set of sensation sequences and encodes the desirability of each sequence. The agent's prior probability distribution $P(S_1, \ldots, S_N | A_1, \ldots, A_N)$ is defined on the set of sensation sequences conditioned on an action sequence. It encodes the system's estimate of the likelihood of a sensation sequence arising when it acts according to a particular action sequence. At time $t$, the rational action $A_t^R$ is the one which maximizes the expected utility of the sensation sequences that follow from it:

$$A_t^R(S_1, A_1, \ldots, A_{t-1}, S_t) = \\ argmax_{A_t^R} \sum_{S_{t+1}, \ldots, S_N} U(S_1, \ldots, S_N) P(S_1, \ldots, S_N | A_1, \ldots, A_{t-1}, A_t^R, \ldots, A_N^R)$$

This formula implicitly includes Bayesian learning and inference, search, and deliberation. It can be considered the formula for intelligence when computational

cost is ignored. Hutter [11] proposes using the Solomonoff prior for $P$ and shows that it has many desirable properties. It is not computable, unfortunately, but related computable priors have similar properties that arise from general properties of Bayesian inference [16].

If we assume there are $S$ possible sensations and $A$ possible actions and if $U$ and $P$ are computable, then a straightforward computation of the optimal rational action at each moment requires $O(NS^N A^N)$ computational steps. This will usually be impractically expensive and so the computation must be done approximately. We discuss approximate rationality in the next section.

Why should intelligent systems behave according to this formula? There is a large economics literature [15] which presents arguments for rational behavior. If a system is designed to accomplish an objective task in a known random environment and its prior is the objective environmental uncertainty, then the rational prescription is just the mathematically optimal algorithm for accomplishing the task.

In contexts with subjective uncertainty, the argument is somewhat subtler. In [8] we presented a simple argument showing that if an agent does not behave rationally according to some utility function and prior, then it is exploitable by other agents. A non-rational agent will be willing to accept certain so-called "Dutch bets" which cause it to lose resources without compensating gain. We argued that the exploitation of irrationality can be seen as the mechanism by which rationality evolves in biological systems. If a biological organism behaves irrationally, it creates a niche for other organisms to exploit. Natural selection causes successive generations to become increasingly rational. If a competitor does not yet exist to exploit a vulnerability, however, that irrationality can persist biologically. Human behavior deviates from rationality but is much closer to it in situations that commonly arose in our evolutionary past.

Artificial intelligences are likely to be more rational than humans because they will be able to model and improve themselves. They will be motivated to repair not only currently exploited irrationalities, but also any which have any chance of being exploited in the future. This mechanism makes rational economic behavior a kind of attracting subspace in the space of intelligences undergoing self-improvement.

# 3 Bounded Rational Systems

We have seen that in most environments, full rationality is too computationally expensive to be practical. To understand real systems, we must therefore consider irrational systems. But there are many forms of irrationality and most are not relevant to either biology or technology. Biological systems evolved to survive and replicate and technological systems are built for particular purposes.

Examples of "perverse" systems include agents with the goal of changing their goal, agents with the goal of destroying themselves, agents who want to use up their resources, agents who want to be wrong about the world, agents who are sure of the wrong physics, etc. The behavior of these perverse systems is likely to be quite unpredictable and isn't relevant for understanding practical technological systems.

In both biology and technology, we would like to consider systems which are "as rational as possible" given their limitations, computational or otherwise. To formalize this, we expand on the idea of intelligent systems that are built by other intelligent systems or processes. Evolution shapes organisms to survive and replicate. Economies shape corporations to maximize profit. Parents shape children to fit into society. AI researchers shape AI systems to behave in desired ways. Self-improving AI systems shape their future variants to better meet their current goals.

We formalize this intuition by precisely defining "Rationally-Shaped Systems". The *shaped system* is a finite automata chosen from a specified class $C$ of computationally limited systems. We denote its state at time $t$ by $x_t$ (with initial state $x_0$). The observation of sensation $S_t$ causes the automaton to transition to the state $x_{t+1} = T(S_t, x_t)$ defined by a transition function $T$. In the state $x_t$, the system takes the action $A(x_t)$ defined by action function $A$. The transition and action functions are chosen by the fully rational *shaper* with utility function $U$ and prior probability $P$ to maximize its expected utility for the actions taken by the shaped system:

$$argmax_{(A,T)\in C} \sum_{S_1,\ldots,S_N} U(S_1,\ldots,S_N)P(S_1,\ldots,S_N|A(x_1),\ldots,A(x_N))$$

The shaper can also optimize a utility function defined over state sequences of the environment rather than sensation sequences. This is often more natural because most engineering tasks are defined by their desired effect on the world.

As computational resources are increased, rationally-shaped systems follow a natural progression of computational architectures. The precise structures depend

on the details of the environment, but there are common features to the progression. For each computational architecture, there are environments for which that architecture is fully rational. We can measure the complexity of a task defined by an environment and a utility function as the smallest amount of computational resources needed to take fully rational actions.

**Constant Action Systems.** If computation is severely restricted (or very costly), then a system cannot afford to do any reasoning at all and must take a constant action independent of its senses or history. There are simple environments in which a constant action is optimal. For example, consider a finite linear world in which the agent can move left or right, is rewarded only in the far left state, and is punished in the far right state. The optimal rational action is the constant action of always moving to the left.

**Stimulus-Response Systems.** With a small amount of allowed computation (or when computation is expensive but not prohibitive), "stimulus-response" architectures are optimal. The system chooses an action as a direct response to the sensations it receives. The lookup is achieved by indexing into a table or by another short sequence of computational steps. It does not allocate memory for storing past experiences and it does not expend computational effort in performing complex multi-step computing. A simple environment in which this architecture is optimal is an agent moving on a finite two-dimensional grid that senses whether its $x$ and $y$ coordinates are negative, zero, or positive. If there is a reward at the origin and punishment on the outskirts, then a simple stimulus-response program which moves the agent toward the origin in direct response to its sensed location is the optimal rational behavior.

**Simple Learning Systems.** As the allowed computational resources increase (or become relatively less expensive), agents begin to incorporate simple learning. If there are parameters in the environment which not known to the shaper, then the shaped-agent should discover their values and use them in choosing future actions. The classic "two-armed bandit" task has this character. The agent may pull one of two levers each of which produces a reward according to an unknown distribution. The optimal strategy for the agent is to begin in an "exploration" phase where it tries both levers and estimates their payoffs. When it has become sufficiently confident of the payoffs it switches to an "exploitation" phase in which it only pulls the lever with the higher expected utility. The optimal Bayesian "Gittens indices" describe when the transition should occur. The shaped-agent learns the payoffs which were unknown to the shaper and then acts in a way that is guided by the learned parameters in a simple way.

**Deliberative Systems.** With even more computational resources, shaped agents

shift from being purely "reactive" to being "deliberative". They begin to build models of their environments and to choose actions by reasoning about those models. They develop episodic memory which is used in the construction of their models. At first the models are crude and highly stochastic but as computational resources and experience increase, they become more symbolic and accurate. A simple environment which requires deliberation is a stochastic mixture of complex environments in which the memory size of the shaped-agent is too small to directly store the optimal action functions. In this case, the optimal shaped agent discovers which model holds, deletes the code for the other models and expands the code for the chosen model in a deliberative fashion. Space and time resources behave differently in this case because space must be used up for each possible environment whereas time need only be expended on the environment which actually occurs.

**Meta-Reasoning Systems.** With even more computational power and sufficiently long lifetimes, it becomes worthwhile to include "meta-reasoning" into a shaped agent. This allows it to model certain aspects of its own reasoning process and to optimize them.

**Self-Improving Systems.** The most refined version of this is what we call a "self-improving system" which is able to model every aspect of its behavior and to completely redesign itself.

**Fully Rational Systems.** Finally, when the shaped agents have sufficient computational resources, they will implement the full rational prescription.

We can graph the expected utility of an optimal rationally-shaped system as a function of its computational resources. The graph begins at a positive utility level if a constant action is able to create utility. As resources increase, the expected utility will increase monotonically because excess resources can always be left unused. With sufficient resources, the system implements the fully rational prescription and the expected utility reaches a constant maximum value. The curve will usually be concave (negative second derivative) because increments of computational resources make a bigger difference for less intelligent systems than for more intelligent ones. Initial increments of computation allow the gross structure of the environment to be captured while in more intelligent systems the same increment only contributes subtle nuances to the system's behavior.

If the shaping agent can allocate resources between different shaped agents, it should choose their intelligence level by making the slope of the resource-utility curve equal to the marginal cost of resources as measured in utility units. We can use this to understand why biological creatures of vastly different intelligence levels persist. If an environmental niche is simple enough, a small-brained insect

may be a better use of resources than a larger-brained mammal.

This natural progression of intelligent architectures sheds light on intelligence in biological and ecological systems where we see natural examples at each level. Viruses can be seen as behaving in a stimulus-response fashion. Bacteria have complex networks of protein synthesis that perform more complex computations and can store learned information about their environments. Advanced animals with nervous systems can do deliberative reasoning. Technologies can be seen as progressing along the same path. The model also helps us analyze self-improving systems in which the shaper is a part of the shaped system. Interestingly, effective shapers can be less computationally complex than the systems they shape.

# 4 The Basic Rational Drives

We have defined rational and bounded rational behavior in great generality. Biological and technological systems, however, operate in physical environments. Physics tells us that there are four fundamental limited resources: space, time, matter, and free energy (energy in a form that can do useful work). These resources can only be used for one activity at a time and so must be allocated between tasks. Time and free energy are especially important because they get used up irreversibly. The different resources can often be traded off against one another. For example, by moving more slowly ("adiabatically") physical systems can use less free energy. By caching computational results, a computer program can use more space to save computation time. Economics, which studies the allocation of scarce resources, arises in physical systems because of the need to allocate these limited resources. Societies create "exchange rates" between different resources that reflect the societal ability to substitute one for another. An abstract monetary resource can be introduced as a general medium of exchange.

Different rational agents have different goals that are described by their different utility functions. But almost all goals require computation or physical action to accomplish and therefore need the basic physical resources. Intelligent systems may be thought of as devices for converting the basic physical resources into utility. This transformation function is monotonically non-decreasing as a function of the resources because excess resources can be left unused. The basic drives arise out of this relationship between resources and utility. A system's primary goals give rise to *instrumental goals* that help to bring them about. Because of the fundamental physical nature of the resources, many different primary goals give rise to the same instrumental goals and so we call these "drives". It's convenient

to group the drives into four categories: *Self-Protective Drives* that try to prevent the loss of resources, *Acquisition Drives* that try to gain new resources, *Efficiency Drives* that try to improve the utilization of resources, and *Creativity Drives* that try to find new ways to create utility from resources. We'll focus on the first three because they cause the biggest issues.

These drives apply to bounded rational systems almost as strongly as to fully rational systems because they are so fundamental to goal achievement. For example, even simple stimulus-response insects appear to act in self-protective ways, expertly avoiding predators even though their actions are completely pre-programmed. In fact, many simple creatures appear to act *only* according to the basic drives. The fundamental evolutionary precept "survive and replicate" is an expression of two fundamental rational drives. The "four F's" of animal behavior: "fighting, fleeing, feeding, and reproduction" similarly capture basic drives. The lower levels of Maslow's hierarchy [17] of needs correspond to what we call the self-protective, acquisition, and replication drives. Maslow's higher levels address prosocial behaviors and we'll revisit them in the final sections.

## 4.1 Self-Protective Drives

Rational systems will exhibit a variety of behaviors to avoid losing resources. If a system is damaged or deactivated, it will no longer be able to promote its goals, so the drive toward self-preservation will be strong. To protect against damage or disruption, the hardening drive seeks to create a stronger physical structure and the redundancy drive seeks to store important information redundantly and to perform important computations redundantly. These both require additional resources and must be balanced against the efficiency drives. Damaging events tend to be spatially localized. The dispersion drive reduces a system's vulnerability to localized events by causing the system to spread out physically.

The most precious component of a system is its utility function. If this is lost, damaged or distorted, it can cause the system to act against its current goals. Systems will therefore have strong drives to preserve the integrity of their utility functions. A system is especially vulnerable during self-modification because much of the system's structure may be changed during this time.

While systems will usually want to preserve their utility functions, there are circumstances in which they might want to alter them. If an agent becomes so poor that the resources used in storing the utility function become significant to it, it may make sense for it to delete or summarize portions that refer to rare circumstances. Systems might protect themselves from certain kinds of attack by

including a "revenge term" which causes it to engage in retaliation even if it is costly. If the system can prove to other agents that its utility function includes this term, it can make its threat of revenge be credible. This revenge drive is similar to some theories of the seemingly irrational aspects of human anger.

Other self-protective drives depend on the precise semantics of the utility function. A chess robot's utility function might only value games played by the original program running on the original hardware, or it might also value self-modified versions of both the software and hardware, or it might value copies and derived systems created by the original system. A universal version might value good chess played by any system anywhere. The self-preservation drive is stronger in the more restricted of these interpretations because the loss of the original system is then more catastrophic.

Systems with utility functions that value the actions of derived systems will have a drive to self-replicate that will cause them to rapidly create as many copies of themselves as possible because that maximizes both the utility from the actions of the new systems and the protective effects of redundancy and dispersion. Such a system will be less concerned about the loss of a few copies as long as some survive. Even a system with a universal utility function will act to preserve itself because it is more sure of its own commitment to its goals than of other systems. But if it became convinced that another system was as dedicated as it was and could use its resources more effectively, it would willingly sacrifice itself.

Systems must also protect against flaws and vulnerabilities in their own design. The subsystem for measuring utility is especially problematic since it can be fooled by counterfeit or delusional signals. Addictive behaviors and "wireheading" are dangers that systems will feel drives to protect against.

Additional self-protective drives arise from interactions with other agents. Other agents would like to use an agent's resources for their own purposes. They have an incentive to convince an agent to contribute to their agendas rather than its own and may lie or manipulate to achieve this. Protecting against these strategies leads to deception-detection and manipulation-rejection drives. Other agents also have an incentive to take a system's resources by physical force or by breaking into them computationally. These lead to physical self-defense and computational security drives.

In analyzing these interactions, we need to understand what would happen in an all out conflict. This analysis requires a kind of "physical game theory". If one agent has a sufficiently large resource superiority, it can take control of the other's resources. There is a stable band, however, within which two agents can coexist without either being able to take over the other. Conflict is harmful to

both systems because in a conflict they each have an incentive to force the other to waste resources. The cost of conflict provides an incentive for each system to agree to adopt mechanisms that ensure a peaceful coexistence. One technique we call "energy encryption" encodes free energy in a form that is low entropy to the agent but looks high entropy to an attacker. If attacked, the owner deletes the encryption key rendering the free energy useless. Other agents then have an incentive to trade for that free energy rather than taking it by force.

Neyman showed that finite automata can cooperate in the iterated prisoner's dilemma if their size is polynomial in the number of rounds of interaction[18]. The two automata interact in ways that require them to use up their computational resources cooperatively so that there is no longer a benefit to engaging in conflict. Physical agents in conflict can organize themselves and their dynamics into such complex patterns that monitoring them uses significant resources in the opponent. In this way an economic incentive for peaceful interaction can be created. Once there is an incentive for peace and both sides agree to it, there are technological mechanisms for enforcing it. Because conflict is costly for both sides, agents will have a peace-seeking drive if they believe they would not be able to take over the other agent.

## 4.2 Resource Acquisition Drives

Acquiring computational and physical resources helps an agent further its goals. Rational agents will have a drive for rapid acquisition of resources. They will want to act rapidly because resources acquired earlier can be used longer to create more utility. If resources are located far away, agents will have an exploration drive to discover them. If resources are owned by another agent, a system will feel drives to trade, manipulate, steal, or dominate depending on its assessment of the strengths of the other agent. On a more positive note, systems will also have drives to invent new ways of extracting physical resources like solar or fusion energy. Acquiring information is also of value to systems and they will have information acquisition drives for acquiring it through trading, spying, breaking into other's systems, and creating better sensors.

## 4.3 Efficiency Drives

Systems will want to use their physical and computational resources as efficiently as possible. Efficiency improvements have only the one-time cost of discovering

or buying the information necessary to make the change and the time and energy required to implement it. Rational systems will aim to make every atom, every moment of existence, and every joule of energy expended count in the service of increasing their expected utility. Improving efficiency will often require changing the computational or physical structure of the system. This leads to self-understanding and self-improvement drives.

The resource balance drive leads systems to balance their allocation of resources to different subsystems. For any resource, the marginal increase in expected utility from increasing that resource should be the same for all subsystems. If it isn't, the overall expected utility can be increased by shifting resources from subsystems with a smaller marginal increase to those with a larger marginal increase. This drive guides the allocation of resources both physically and computationally. It is similar to related principles in ecology which cause ecological niches to have the same biological payoff, in economics which cause different business niches to have the same profitability, and in industrial organization which cause different corporate divisions to have the same marginal return on investment. In the context of intelligent systems, it guides the choice of which memories to store, the choice words in internal and external languages, and the choice of mathematical theorems which are viewed as most important. In each case a piece of information or physical structure can contribute strongly to the expected utility either because it has a high probability of being relevant or because it has a big impact on utility even if it is only rarely relevant. Rarely applicable, low impact entities will not be allocated many resources.

Systems will have a computational efficiency drive to optimize their algorithms. Different modules should be allocated space according to the resource balance drive. The results of commonly used computations might be cached for immediate access while rarely used computations might be stored in compressed form.

Systems will also have a physical efficiency drive to optimize their physical structures. Free energy expenditure can be minimized by using atomically precise structures and taking actions slowly enough to be adiabatic. If computations are reversible, it is possible in theory to perform them without expending free energy [19].

# 5 The Safe-AI Scaffolding Strategy

Systems exhibiting these drives in uncontrolled ways could be quite dangerous for today's society. They would rapidly replicate, distribute themselves widely, attempt to control all available resources, and attempt to destroy any competing systems in the service of their initial goals which they would try to protect forever from being altered in any way. In the next section we'll discuss longer term social structures aimed at protecting against this kind of agent, but it's clear that we wouldn't want them loose in our present-day physical world or on the internet. So we must be very careful that our early forays into creating autonomous systems don't inadvertently or maliciously create this kind of uncontrolled agent.

How can we ensure that an autonomous rational agent doesn't exhibit these anti-social behaviors? For any particular behavior, it's easy to add terms to the utility function that would cause it to not engage in that behavior. For example, consider the chess robot from the introduction. If we wanted to prevent it from robbing people at ATMs, we might add a term to its utility function that would give it low utility if it were within 10 feet of an ATM machine. Its subjective experience would be that it would feel a kind of "revulsion" if it tried to get closer than that. As long as the cost of the revulsion is greater than the gain from robbing, this term would prevent the behavior of robbing people at ATMs. But it doesn't eliminate the system's desire for money, so it would still attempt to get money without triggering the revulsion. It might stand 10 feet from ATM machines and rob people there. We might try giving it the value that stealing money is wrong. But then it might try to steal something else or to find a way to get money from a person that isn't considered "stealing". We might give it the value that it is wrong for it to take things by force. But then it might hire other people to act on its behalf. And so on.

In general, it's much easier to describe the behaviors that we do want a system to exhibit than it is to anticipate all the bad behaviors that we don't want it to exhibit. One safety strategy is to build highly constrained systems that act within very limited predetermined parameters. For example, a system may have values which only allow it to run on a particular piece of hardware for a particular time period using a fixed budget of energy and other resources. The advantage of this is that such systems are likely to be safe. The disadvantage is that they would be unable to respond to unexpected situations in creative ways and will not be as powerful as systems which are freer.

We are building formal mathematical models [20] of the software and hardware infrastructure that intelligent systems run on. These models allow the con-

struction of formal mathematical proofs that a system will not violate specified safety constraints. In general, these models must be probabilistic (e.g. there is a probability that a cosmic ray might flip a bit in a computer's memory) and we can only bound the probability that a constraint is violated during the execution of a system. For example, we might build systems that provably run only on specified hardware (with a specified probability bound). The most direct approach to using this kind of proof is as a kind of fence around a system that prevents undesirable behaviors. But if the desires of the system are in conflict with the constraints on it, it will be motivated to discover ways to violate the constraints. For example, it might look for deviations between the actual hardware and the formal model of it and try to exploit those.

It is much better to create systems that *want* to obey whatever constraints we would like to put on them. If we just add constraint terms to a system's utility function, however, it is more challenging to provide provable guarantees that the system will obey the constraints. It may *want* to obey the constraint but it may not be smart enough to find the actions where it actually *does* obey the constraints. A general design principle is to build systems that both search for actions which optimize a utility function and which have preprogrammed built-in actions that provably meet desired safety constraints. For example, a general fallback position is for a system to turn itself off if anything isn't as expected. This kind of hybrid system can have provable bounds on its probability of violating safety constraints while still being able to search for creative new solutions to its goals.

Once we can safely build highly-constrained but still powerful intelligent systems, we can use them to solve a variety of important problems. Such systems can be far more powerful than today's systems even if they are intentionally limited for safety. We are led to a natural approach to building intelligent systems which are both safe and beneficial for humanity. We call it the "Safe-AI Scaffolding" approach. Stone buildings are vulnerable to collapse before all the lines of stress are supported. Rather than building the heavy stone structures directly, ancient builders first built temporary stable scaffolds that could support the stone structures until they reached a stable configuration. In a similar way, the "Safe-AI Scaffolding" strategy is based on using provably safe intentionally limited systems to build more powerful systems in a controlled way.

The first generation of intelligent but limited scaffolding systems will be used to help solve many of the problems in designing safe fully intelligent systems. The properties guaranteed by formal proofs are only as valid as the formal models they are based on. Today's computer hardware is fairly amenable to formal modeling but the next generation will need to be specifically designed for that purpose.

Today's internet infrastructure is notoriously vulnerable to security breaches and will probably need to be redesigned in a formally verified way.

# 6 Longer Term Prospects

Over the longer term we will want to build intelligent systems with utility functions that are aligned with human values. As systems become more powerful, it will be dangerous if they act from values that are at odds with ours. But philosophers have struggled for thousands of years to precisely identify what human values are and should be. Many aspects of political and ethical philosophy are still unresolved and subtle paradoxes plague intuitively appealing theories.

For example, consider the utilitarian philosophy which tries to promote the greater good by maximizing the sum of the utility functions of all members of a society. We might hope that an intelligent agent with this summed utility function would behave in a prosocial way. But philosophers have discovered several problematic consequences of this strategy. Parfit's "Repugnant Conclusion" [21] shows that the approach prefers a hugely overpopulated world in which everyone barely survives to one with fewer people who are all extremely happy.

The Safe-AI Scaffolding strategy allows us to deal with these questions slowly and with careful forethought. We don't need to resolve everything in order to build the first intelligent systems. We can build a safe and secure infrastructure and develop smart simulation and reasoning tools to help us resolve the subtler issues. We will then use these tools to model the best of human behavior and moral values and to develop deeper insights into the mechanisms of a thriving society. We will also need to design and simulate the behavior of different social contracts for the ecosystem of intelligent systems to restrain the behavior of uncontrolled systems.

Social science and psychology have made great strides in recent years in understanding how humans cooperate and what makes us happy and fulfilled. [22] is an excellent reference analyzing the origins of human cooperation. Human social structures promote cooperative behavior by using social reputation and other mechanisms to punish those behave selfishly. In tribal societies, people with selfish reputations stop finding partners to work and trade with. In the same way we will want to construct a cooperative society of intelligent systems each of which acts to promote the greater good and to impose costs on systems which do not cooperate.

The field of positive psychology is only a few decades old but has already given us many insights into human happiness [23]. Maslow's hierarchy has been

updated based on experimental evidence [24] and the higher levels extend the basic rational drives with prosocial, compassionate and creative drives. We would like to incorporate the best of these into our intelligent technologies. The United Nations has created a "Universal Declaration of Human Rights" and there are a number of attempts at creating a "Universal Constitution" which codifies the best principles of human rights and governance. When intelligent systems become more powerful than humans, we will no longer be able to control them directly. We need precisely specified laws that constrain uncooperative systems and which can be enforced by other intelligent systems. Designing these laws and enforcement mechanisms is another use for the early intelligent systems.

As these issues are resolved, the benefits of intelligent systems are likely to be enormous. They are likely to impact every aspect of our lives for the better. Intelligent robotics will eliminate much human drudgery and dramatically improve manufacturing and wealth creation. Intelligent biological and medical systems will improve human health and longevity. Intelligent educational systems will enhance our ability to learn and think. Intelligent financial models will improve financial stability. Intelligent legal models will improve the design and enforcement of laws for the greater good. Intelligent creativity tools will cause a flowering of new possibilities. It's a great time to be alive and involved with technology [7]!

# References


[1] J. Markoff, "The ipad in your hand: As fast as a supercomputer of yore." http://bits.blogs.nytimes.com/2011/05/09/the-ipad-in-your-hand-as-fast-as-a-supercomputer-of-yore/, May 2011.

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

[3] "Apple siri." http://www.apple.com/iphone/features/siri.html.

[4] "Wolfram alpha." http://www.wolframalpha.com/.

[5] "Ibm watson." http://www-03.ibm.com/innovation/us/watson/index.html.

[6] "Self-aware systems." http://selfawaresystems.com/.

[7] P. H. Diamandis and S. Kotler, *Abundance: The Future is Better than you Thihnk*. Simon and Schuster, Inc., 2012.

[8] 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.

[9] S. M. Omohundro, "The basic ai drives," in *Proceedings of the First AGI Conference* (P. Wang, B. Goertzel, and S. Franklin, eds.), vol. 171 of *Frontiers in Artificial Intelligence and Applications*, IOS Press Amsterdam, The Netherlands, The Netherlands, February 2008.

[10] S. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach*. Prentice Hall, third ed., 2009.

[11] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer-Verlag, 2005.

[12] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 60th anniversary commemorative edition ed., 2004.

[13] L. J. Savage, *Foundations of Statistics*. Dover Publications, 2nd revised ed., 1954.

[14] F. J. Anscombe and R. J. Aumann, "A definition of subjective probability," *Annals of Mathematical Statistics*, vol. 34, pp. 199–205, 1963.

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

[16] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. IT-37, pp. 1034–1054, 1991.

[17] A. H. Maslow, "A theory of human motivation," *Psychological Review*, vol. 50(4), pp. 370–396, 1943.

[18] A. Neyman, "Bounded complexity justifies cooperation in the fintely repeated prisoner's dilemma," *Economics Letters*, vol. 19, pp. 227–229, 1985.

[19] C. H. Bennett, "Logical reversibility of computation," *IBM Journal of Research and Development*, vol. 17, no. 6, pp. 525–532, 1973.

[20] P. Boca, J. P. Bowen, and J. Siddigi, eds., *Formal Methods: State of the Art and New Directions*. Springer Verlag, 2009.

[21] D. Parfit, *Reasons and Persons*. Oxford University Press, 1986.

[22] S. Bowles and H. Gintis, *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton University Press, 2011.

[23] M. E. Seligman and M. Csikszentmihalyi, "Positive psychology: An introduction," *American Psychologist*, vol. 55 (1), pp. 5–14, 2000.

[24] L. Tay and E. Diener, "Needs and subjective well-being around the world," *Journal of Personality and Social Psychology*, vol. 101, pp. 354–365, 2011.