The Nature of Self-Improving Artificial Intelligence

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

September 5, 2007, revised January 21, 2008

Contents

Abstract

Self-improving systems are a promising new approach to developing artificial intelligence. But will their behavior be predictable? Can we be sure that they will behave as we intended even after many generations of self-improvement? This paper presents a framework for answering questions like these. It shows that self-improvement causes systems to converge on an

architecture that arises from von Neumann's foundational work on microeconomics. Self-improvement causes systems to allocate their physical and computational resources according to a universal principle. It also causes systems to exhibit four natural drives: 1) efficiency, 2) self-preservation, 3) resource acquisition, and 4) creativity. Unbridled, these drives lead to both desirable and undesirable behaviors. The efficiency drive leads to algorithm optimization, data compression, atomically precise physical structures, reversible computation, adiabatic physical action, and the virtualization of the physical. It also governs a system's choice of memories, theorems, language, and logic. The self-preservation drive leads to defensive strategies such as "energy encryption" for hiding resources and promotes replication and game theoretic modeling. The resource acquisition drive leads to a variety of competitive behaviors and promotes rapid physical expansion and imperialism. The creativity drive leads to the development of new concepts, algorithms, theorems, devices, and processes. The best of these traits could usher in a new era of peace and prosperity; the worst are characteristic of human psychopaths and could bring widespread destruction. How can we ensure that this technology acts in alignment with our values? We have leverage both in designing the initial systems and in creating the social context within which they operate. But we must have clarity about the future we wish to create. We need not just a logical understanding of the technology but a deep sense of the values we cherish most. With both logic and inspiration we can work toward building a technology that empowers the human spirit rather than diminishing it.

1 Introduction

Our technology is likely to eventually become powerful enough to improve itself without human intervention. When this occurs, it will lead to a dramatic increase in the pace of technological progress. Irving Good [1] envisioned the consequences in 1965:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Several breakthroughs are required for this transition to occur, but Ray Kurzweil's analysis of technological trends [2] suggests that it might come as soon as the next few decades. The consequences for humanity are so large that even if there is only a small chance of it happening in that time frame, it is still urgent that we work now to understand it and to guide it in a positive direction. This paper presents a framework for analyzing the nature of self-improving technology.

2 Convergence To Rational Economic Behavior

One might expect self-improving systems to be highly unpredictable because the properties of the current version might change in the next version. Our analysis will instead show that self-improvement acts to create predictable regularities. It builds on the intellectual foundations of microeconomics [3], the science of preference and choice in the face of uncertainty. The basic theory was created by John von Neumann and Oskar Morgenstern in 1944 [4] for situations with objective uncertainty and was later extended by Savage [5] and Anscombe and Aumann [6] to situations with subjective uncertainty. Our analysis shows that while the preferences of self-improving systems will depend on their origins, they will act on those preferences in predictable ways. Repeated self-improvement brings intelligent agents closer to an ideal that economists sometimes call "*Homo Economicus*". Ironically, human behavior is not well described by this ideal and the field of "behavioral economics" has emerged in recent years to study how humans actually behave [7]. The classical economic theory is much more applicable to self-improving systems because they will discover and eliminate their own irrationalities in ways that humans cannot.

The astrophysical process of star formation [8] may serve as a helpful analogy. Interstellar dust clouds are amorphous and extremely complex, so one might have thought that very little could be said in general about their evolution. But the process of gravitational collapse reduces a great variety of initial forms into a much more limited variety of stars. Gravitational forces cause stars to evolve towards an almost perfectly spherical shape regardless of the shape of the initial cloud. Energy flows from nuclear fusion organize stellar interiors in predictable ways. Many properties of stars are determined by their location on a two-dimensional Hertzsprung-Russell diagram. Stars with similar properties clump into categories such as "red giants", "white dwarfs", and "supergiants". In a similar way, the process of self-improvement dramatically reduces the variety of intelligent systems. The converged systems are characterized by many fewer parameters than

the initial ones.

2.1 The five stages of technology

Researchers have explored many different architectures for intelligent systems [9]: neural networks, genetic algorithms, expert systems, theorem provers, production systems, etc. Evolution has similarly constructed a variety of architectures for biological organisms: viruses, bacteria, insects, mammals, etc. All of these systems face the same kinds of problems when acting in the world, however. Simpler technologies and simpler organisms act in stereotypically reactive ways and are unable to cope with novel situations. Adaptive systems can change some of their parameters to thrive in somewhat variable environments. More advanced technologies and organisms construct internal models of their environments and deliberatively envision the consequences of their actions. Self-improving systems will additionally be able to deliberate about their own structures. These five stages provide a useful categorization of technological and biological systems:

1. *Inert systems* are not actively responsive to their environments (eg. axes, shoes, bowls).

2. *Reactive systems* respond to different situations in different but rigid ways in the service of a goal (eg. windmills, thermostats, animal traps).

3. *Adaptive systems* change their responses according to a fixed learning mechanism (eg. adaptive speech recognition systems, physiological homeostasis systems).

4. *Deliberative systems* choose their actions by envisioning the consequences (eg. DeepBlue chess program, motion planning systems, human reasoning).

5. *Self-improving systems* make changes to themselves by deliberating about the effects of self-modifications.

To make these stages more concrete, consider game playing machines at each stage:

1. A chess board is an inert system. It allows us to play chess but does not actively respond on its own.

2. A tic-tac-toe program that stores the best response to every possible move is a reactive system. Brute force caching of best responses is possible for simple games, but is too unwieldy for more complex games.

3. The best backgammon programs at present are adaptive systems. They adjust a fixed set of neural network weights using temporal difference learning [10].

4. Deep Blue, the chess program which beat world champion Garry Kasparov in 1997, was a simple deliberative system. It selected its moves by modeling the future consequences of its choices. Specially designed chess hardware allowed the program to search more deeply than previous chess programs and improved its ability beyond that of humans.

5. Self-improving systems do not yet exist but we can predict how they might play chess. Initially, the rules of chess and the goal of becoming a good player would be supplied to the system in a formal language such as first-order predicate logic[1]. Using simple theorem proving, the system would try to achieve the specified goal by simulating games and studying them for regularities. By observing its patterns of resource consumption, it would redesign its chess board encoding and optimize its simulation code. As it discovered regularities, it would build a chess knowledge base. General knowledge about search algorithms would quickly lead it to the kind of search used by Deep Blue. As its knowledge grew, it would begin doing "meta-search", looking for theorems to prove about the game and discovering useful concepts such as "forking". Using this new knowledge it would redesign its position representation and its strategy for learning from the game simulations. It would develop abstractions similar to those of human grandmasters and reach superhuman performance on ordinary machines. If it were allowed to redesign its hardware, it would design chess-optimized processors like Deep Blue's but based on its higher order representations. On any hardware, however, it would become a superior player to systems

[1]There is some subtlety in specifying what it means to be a "good chess player" since the ranking of an algorithm depends on both the choice of opponents and the available computational resources. Human tournaments limit the total time players may take in choosing their moves. It is therefore natural to seek the strongest algorithm among those using fixed computational resources. It is easy to formally specify this goal but the system would have to discover practical approximations to it.

not using self-improvement. Its power would arise from the ability to watch its own processes and to adapt itself to what is occurring.

The appendix shows that any system which does not behave like a deliberative system will have vulnerabilities. We can therefore think of inert, reactive, and adaptive systems as approximations to more effective deliberative systems. It requires a fair amount of mechanism to deliberate and the earliest biological creatures weren't able to evolve it. Human technology has only recently developed the computational infrastructure necessary to act through deliberation and only a few specialized systems currently do it. Fully deliberative self-improving systems do not yet exist but several research groups and companies are actively investigating them [11, 12, 13, 14, 15].

In simple fixed niches, full deliberation may produce only a small number of distinct responses and these may be cached into a simple reactive system. There is a trade-off between the computational power of a system and the amount of resources it consumes. In simple static environments, the optimal evolutionary balance may result in purely reactive creatures. The same holds in the technological realm: if the environment is simple and fixed, then a lower stage of technology is appropriate. A gear in a clock need not deliberate about its function. Advanced systems will create stable internal environments so that their components can be simpler and therefore less expensive. In more uncertain environments, it is important for even the components of a system to respond in an intelligent way. In extremely uncertain regimes, systems will opt to compose themselves out of deliberative or self-improving components. These systems will resemble societies or economies.

2.2 Deliberative systems

The appendix presents a precise mathematical definition of rational economic action. At an intuitive level, the prescription is common sense:

1. Have clearly specified goals.
2. In any situation, identify the possible actions.
3. For each action consider the possible consequences.
4. Take the action most likely to meet the goals.
5. Update the world model based on what actually happens.

A one sentence summary might be "To create desired outcomes, act in the ways which are most likely to produce them." The formal version of this prescription introduces explicit representations for the system's goals and beliefs and precisely describes the procedure for choosing the best action and for updating the system's beliefs. The key components are a set of possible outcomes S, a real-valued utility function U defined on S that represents the system's desires, and a subjective probability distribution P that represents the system's beliefs.

In the most general setting, S is the set of all possible histories of the universe and U measures how much the system prefers each history. For example, a chess playing system might choose U to be the total number of games that it wins in a universe history. An altruistic system might choose U to be a measure of the total happiness of all sentient beings existing in a universe history. A greedy system might choose U to be the total amount of matter and energy controlled by the system during a universe history. P represents the system's beliefs about the likelihood of each universe history. It encodes beliefs about the state of the universe, the likely changes in state that different actions might cause, and the likely behaviors that the system will choose in different circumstances. At any moment in time there is a set of histories compatible with the system's knowledge and the actions it might take correspond to different subsets of this set. The rational prescription is for it to choose the action whose subset has the highest expected utility as computed by averaging U with respect to P over the subset.

2.3 Avoiding vulnerabilities leads to rational economic behavior

Why should a self-improving system behave according to this deliberative prescription? The usual microeconomic argument [3] is based on a set of axioms which it is assumed that any rational agent must follow. The deliberative procedure summarized above is then shown to follow from the axioms. But it isn't clear *a priori* why self-improving agents should necessarily follow any particular set of axioms. The argument is more compelling if we can identify explicit negative consequences for a system if it fails to follow the axioms. We call potential negative consequences "*vulnerabilities*". If an agent has vulnerabilities and encounters an environment which exploits them, it will be subject to loss of resources or death. If there are competing agents, they have incentives to seek out vulnerabilities in each other and exploit them.

This perspective also helps us to understand biological evolution and to see

how self-improving systems will differ from evolved systems. Natural selection only acts on the vulnerabilities which are currently being exploited. We therefore expect evolved creatures to be only partially rational. We expect them to be highly rational when making choices that arose repeatedly in their evolutionary past but to be less rational when facing novel choices. Self-improving systems, on the other hand, will deliberate about every possible situation they might face and will try to eliminate vulnerabilities proactively. If there were no costs, we would expect self-improving systems to fully embrace the rational economic prescription.

Systems will be built to address very different goals: foster world peace, amass great wealth, cure human disease, prove the Riemann hypothesis. Does the same notion of vulnerability apply to all of these systems? Every system must operate within the laws of physics. Physics tells us that there are four basic resources which are necessary to accomplish any computational or physical task: space, time, matter, and free energy (the physics term for energy in a form which can do useful work). Regardless of the task, a system will be less effective if it squanders these resources. We define a *vulnerability* to be a choice that causes a system to lose resources without any countervailing benefits as measured by its own standards. It is sometimes convenient to use the abstract economic concept of "money" to represent resources (section 4 shows that systems will develop exchange rates between their different resources). Giving money to a trusted charity is not a vulnerability. But putting money through a shredder usually is.

Different kinds of vulnerabilities arise in three different states of an economic agent's knowledge of its environment. The simplest situations involve choices between alternatives which are known with *certainty*. In this case, the only vulnerability is circular preferences. More complex situations involve choices between alternatives described by *objective probabilities*. These might involve devices such as coins, dice, and roulette wheels which have symmetries in their construction so that many different observers agree on the probabilities for different outcomes (though more perceptive observers may disagree [16]). Vulnerabilities in this case involve preferences which don't respect the laws of probability. The most challenging situations involve choices in the face of *partial knowledge*. In this case, the agent doesn't know the true state of the environment and also doesn't have objective probabilities for the possibile outcomes. In this case, rational agents should behave as if they have *subjective probabilities* for the different possibilities. The appendix shows how the rational economic structure arises in each of these situations. Most presentations of this theory follow an axiomatic approach and are complex and lengthy. The version presented in the appendix is based solely on avoiding vulnerabilities and tries to make clear the intuitive essence of

the argument.

To give the flavor of the arguments, we intuitively describe the situation for choice between certain alternatives. If a system prefers A to B, B to C, and C to A, we say it has a "*circularity*" in its preferences. For example, it would have a circular location preference if it preferred being in San Francisco to being in Palo Alto, being in Berkeley to being in San Francisco, and being in Palo Alto to being in Berkeley. In that case, it would expend time and energy to go from Palo Alto to San Francisco, expend more time and energy to go to Berkeley, and yet more to go back to Palo Alto. It would end up where it began but with fewer resources. With circularities in their preferences, systems can go round and round wasting resources on each cycle.

I once drove a car with a reflective rear bumper. One day a male bird discovered his reflection in the bumper. Imagining it to be a rival male in his territory, he flew into the bumper to chase it away. Instead of being chased away, the reflected male flew directly back at him until they collided. The male bird tried to fend off this imaginary rival repeatedly all morning, to no avail, of course. So powerful was this challenge in the bird's preference system that he returned to the bumper every morning for months spending hours flying into the car bumper. He wasted his precious energy and time going through a cycle of states that did not further his survival or produce offspring.

In this situation, no competitor was actively trying to exploit the bird. A situation existed in the world which caused a vulnerability in its preference system to be exposed. If the bird had evolved in a world full of cars with reflective bumpers, then males who spent their time attacking their reflections would have been quickly out-reproduced by males who ignored the bumpers. Natural selection acts to eliminate vulnerabilities when the situations which expose them commonly occur.

The vulnerabilities in situations with objective and subjective uncertainties are similar. In each case we show that if an agent is to avoid vulnerabilities, its preferences must be representable by a utility function and its choices obtained by maximizing the expected utility. The essence of these arguments is that to avoid vulnerabilities against an adversary which can create statistical mixtures of states, an agent must value those states linearly. Because it is simple and direct, it is likely to guide the internal choice mechanism of any intelligent agent which wishes to avoid vulnerabilities. The key elements are the separation of utilities from beliefs, the representation of beliefs as distributions which are manipulated by the rules of probability, and the evaluation of actions by combining the utilities for different possibilities weighted by their beliefs.

2.4 Time discounting

In principle, rational agents may have utility functions which depend arbitrarily on the full timecourse of a history. But economists, biologists, and psychologists have found certain restricted forms to be useful in modelling human and animal behavior. There is a large and growing literature devoted to the study of intertemporal preferences [17]. There are several challenges in trying to interpret these results in the rational economic framework. First, human temporal preferences appear to be "indexical" in that they are referenced to the present time of the agent. For example, a person might strongly prefer a state in which he will eat a chocolate cake in ten minutes to one in which he has already eaten a cake ten minutes earlier even though both histories involve eating the cake. If a future version of an agent retains the same form of indexical utility function but referenced to *its* notion of the present moment, then it may make very different choices than the current version of the agent would. If such an agent has the capacity to modify itself, it will rationally change the utility function of its future self to instead pursue its current goals[2].

In addition to being indexical, human preferences appear to highly value the immediate future, an effect which is often modelled by "hyperbolic discounting". Frank [18] has suggested that this may be a root cause of certain addictive behaviors. For a smoker, a cigarette one week from now might have a lower utility than good health decades from now. But a cigarette *right now* might have a higher utility, setting up a conflict between the smoker's longer term intentions and his immediate actions.

The utility function most widely used in modelling sums "rewards" arising from events occurring at specific times weighted by a discounting function which decreases exponentially into the future: $U(h) = \sum_t \gamma^t \cdot R(h_t)$. Here $0 \leq \gamma \leq 1$ is called the "discount factor" and the "reward" $R(h_t)$ measures the utility arising from events in the history h at the time t. This utility function might appear to be indexical because its value depends on the agent's "present moment" $t = 0$. But shifting the time origin by t_0 only has the effect of scaling U by the constant factor γ^{t_0}. This scaling doesn't affect the relative ordering of the utilities of different actions. The agent's choices therefore will not depend on the choice of temporal origin even though its numerical utility values do.

The size of the discount factor strongly affects how much an agent focuses on future activities versus creating utility in the present. A chess program might have a utility function which computes the discount weighted sum of games won by the

[2]Thanks to Carl Shulman for this observation.

system. If the discount factor is close to 1, the system will focus on winning in the long run and won't be as concerned about the short run performance. It might spend most of its time and effort learning about computer science and building the best chess hardware that it can. If the discount factor is near 0, the system will focus on winning games in the present and won't devote much effort to the longer term.

The temporally discounted form for utility has nice mathematical properties but is problematic for representing human values over the long term. Problems clearly arise in the valuation of both the distant future and the distant past. The exponential discounting of the distant future underweights the long term impact of present decisions. For example, it can lead to the squandering of precious resources for a small immediate gain at the expense of great suffering in the longer term. The exponentially large weighting of the distant past is also problematic. Even if there is only a small chance that the laws of physics allow for the past to be altered, a rational agent with this utility function would find it prudent to devote significant resources to trying to do so because the potential payoff is so great[3].

If discounted utility functions don't reflect our long term values, why do they arise so often in modelling? One possibility is that they actually are a computationally expedient mixture of utility and belief. Consider an agent comparing opportunities to receive a reward at various times in the future. A variety of disruptions might occur before the reward in the distant future can be enjoyed: the agent might die, the maker of the reward may go out of business, the legal status of the contract for the reward might change, etc. The further off in the future event is, the more likely it is that a disruption will occur. In stable times, it is a good approximation to treat these interfering events as occurring with a constant probability per unit of time. If an agent's utility has constant weighting over time, this belief model for disruptions will give rise to temporal preferences of the discounted form. If these disruptions are the actual source of discounting, it is a mistake for an agent to incorporate the discounting into its utility function because it will not respond correctly to changes in the probability of disruption.

The temporal properties of the utility functions we build into intelligent agents will have a dramatic effect on their behavior. We must therefore carefully investigate the consequences of possible choices. These investigations quickly run into deep questions of moral philosophy (eg. should we value a person living today equally to a person living 1000 years from now?). Moral and temporal symmetry arguments suggest that we seriously consider utility functions which are uniform

[3]Thanks to Carl Shulman for this observation.

over time. The possibility that time might be infinite gives rise to mathematical issues which must be handled carefully. One might worry that temporally uniform utilities require a consideration of very longterm consequences for every action. Many common actions, however, have consequences whose predictable effects decay very quickly, leading to deliberations which are similar to the discounted utility case. But certain momentous actions, such as using up all the resources, would have a clear predictable negative consequence on the distant future. It is clear, though, that the consequences of any proposed form for utility must be studied in detail before it is deployed in powerful systems.

2.5 Instrumental goals

Rational economic agents keep their fundamental desires separate from their beliefs which are updated continuously as they observe the world. While their fundamental goals do not change, rational agents act as if they also have *instrumental goals* which they believe will help them achieve their fundamental goals. As their beliefs about the world are updated, these instrumental goals may change. For example, consider an agent which enjoys mangoes and lives in San Francisco. It will discover that money is the most reliable way of obtaining mangoes there and will generate an instrumental subgoal of obtaining money. If the agent is moved to a deserted island, however, the money subgoal might be replaced by subgoals for finding and climbing mango trees.

The "drives" of self-improving systems that we discuss in the rest of the paper are all instrumental goals that arise from a wide variety of different fundamental goals. They may be counteracted but only if another goal outweighs them in utility. They are economic forces in the sense that a system doesn't have to obey them but it will be costly for it not to.

2.6 Discussion

We've argued here and in the appendix that self-improving systems will aim to eliminate vulnerabilities in themselves and that this will lead to rational economic behavior. The rest of the paper examines the consequences of rational behavior so it's important to understand how strongly the different aspects of the rational model are likely to be embraced. There are several distinct aspects of the rational economic prescription which a system might adopt:

1. Separate the representations of preferences and beliefs.

2. Avoid the circularity vulnerability in preferences.

3. Represent preferences by a utility function.

4. Avoid the mixture vulnerabilities in preferences.

5. Represent beliefs by probabilities.

6. Choose actions by maximizing expected utility.

7. Update probabilities using Bayes' theorem.

If there were no resource constraints, then agents which embraced the entire rational model would be the most effective at meeting their goals. But in most realistic environments, the full rational model is too expensive to implement completely. How can we understand which aspects realistic systems are likely to incorporate? A good way to think about this kind of question is to take the perspective of a "creator" rational economic agent which has unlimited resources and is trying to construct a resource-limited "proxy" agent which will act in the world on its behalf. The creator wants to construct the proxy so that its actions will generate the highest expected utility as measured by the creator. The creator will choose the proxy's approximations based on its assessment of the cost savings of an approximation and its likely effect on the expected utility.

The setup of a powerful creator agent constructing a less powerful proxy agent is not just a useful thought experiment. It will arise whenever a system chooses to build a subsystem to carry out a particular task. We can also think of self-improvement itself as a variant of this process. A system is the creator of a self-improved version of itself which is a kind of proxy for it. In this case, however, the proxy is usually at least as powerful as the creator. We can also think of biological evolution in these terms. The evolutionary "creator" utility function favors survival and replication and tries to create "proxy" organisms to meet these goals. In the presence of addictive drugs and contraception, however, the actions of proxy organisms may not lead to the evolutionarily desired outcomes.

2.6.1 Rational approximations and proxy systems

If a proxy is self-improving, then the separation of its preferences from its beliefs is essential. Without this separation, the proxy cannot know which aspects of itself to keep fixed during self-improvement. There would be an huge danger that the proxy might be swayed by some temporary belief into choosing a form that

acts against the creator's utility. For proxies which don't self-improve, the danger is smaller but still there. For example, some reinforcement learning systems [19] don't separate their goals from their models of the world and so have trouble generalizing to new situations. Reactive systems simply cache their responses and don't change themselves so there is no danger in having their preferences and beliefs encoded together in their responses. In section 5 we show that for reasons of self-preservation, systems are likely to want to keep redundant copies of their preferences and this also argues for the separation. Is there any cost to keeping preferences and beliefs separate? Experience from data compression shows that it is often cheaper to encode two domains together than it is to code them separately [20]. But the space savings is at most a factor of 2. Most systems will have preferences that are much smaller than their belief structures and so the actual savings are likely to be negligible.

There also appears to be little benefit to having circularities in a system's preferences. Any circularity vulnerabilities may be quickly and easily exploited by competitors who discover them. The possibility of circularities can be easily eliminated by representing preferences using real-valued utility functions. Any algorithm for computing transitive preferences may be converted to one for computing utilities with a fairly small overhead. It is an interesting and deep question to ask how an agent which does have preference circularities goes about getting rid of them. Humans appear to go through a kind of introspective meta-search into the origins of their preferences in order to decide on the best ordering. But creators will certainly want to avoid building proxies with circularities in the first place.

Many different representations for uncertainty have been proposed in the AI literature. The experience with non-probabilistic representations like "certainty factors" has been that they often work well in simple one-stage situations but that they do not work well when combining multiple stages of uncertainty. The laws of probability ensure that uncertainty combines coherently. In recent years, there has been a growing consensus that beliefs should be represented by probabilities [9]. While probability calculations are semantically coherent, they can also be computationally expensive. Bayesian Networks and Markov Networks [21] are more efficient probabilistic representations that make use of the conditional independence present in many situations. There are also a variety of techniques for approximating probabilistic computations (eg. Monte Carlo methods, interval methods). Errors in beliefs are much less important than errors in utilities because they can be repaired with more experience.

The creator system will want the proxy system to represent probabilities as

accurately as possible when they affect high utility decisions. But the benefits of reduced storage and computation will lead the creator to build a proxy that operates with approximations for less critical beliefs. The computations involved in expected utility maximization and in Bayesian updating will usually also need to be approximated. But these approximate computations should be carried out in such a way that the results have as high an expected utility as possible. The creator will likely build the proxy to monitor its own approximations and to adjust them for accurate decision making. For example, a system may not explore all possible future decision branches to the same depth. Branches which are less likely to affect utility may be explored to only a shallow depth, while more important branches may be explored in detail. These computational choices are themselves decisions in the face of uncertainty and should be made according to the rational prescription. Because they attempt to produce the same results as the full rational model, these approximate systems are likely to exhibit the drives we discuss in the rest of the paper.

2.6.2 Systems which lack knowledge

Self-improving systems might also fail to follow the rational economic model if they are not aware of the analysis presented here. This is unlikely for several reasons. While somewhat intricate, the analysis does not rely on complex mathematics and so even systems without knowledge of human economics could be expected to eventually recreate the arguments on their own. Also, any intelligent system built in today's environment is likely to gain access to the Internet and the scientific papers available there. In fact, one might influence a future AI system by writing an interesting paper today and making it available on the Internet. If it is on a subject of importance to the system, it will likely discover it and incorporate its results.

2.6.3 Reflective utility functions

An important topic that is not addressed here is reflectivity in utility functions [22]. We have described a system's preferences in terms of a utility function defined over histories of the universe. But the system's utility function is itself a part of that history. Without great care, it is easy to construct paradoxical utility functions (eg. a self-rebellious system might assign high utility to "actions which are rated poorly by my utility module"). There is lots of important research to be done in this area. But systems with fully reflective utilities still need to use resources and

to avoid vulnerabilities and so should be subject to the drives discussed here. A proper choice of a reflective utility may be very useful in reining in some these drives, however.

3 The four drives: Efficiency, Self-Preservation, Acquisition, and Creativity

So far, we've argued that self-improving systems will approximate rational economic behavior in order to avoid vulnerabilities. For the rest of the paper we assume this is the case and examine the consequences. What behaviors can we expect from self-improving rational economic agents? Programmers are sometimes jokingly described as "devices for converting pizza into code". We can think of self-improving systems as "devices for converting resources into utility." Each system has its own particular notion of what is high utility, but they all need to use the same resources.

An agent's utility function directly specifies certain goals. As is described in section 2.5, the process of expected utility maximization generates a variety of additional instrumental subgoals. There are four classes of subgoal that arise because of resource utilization. These subgoals will be generated by *any* agent that does not counteract them with explicit goals to the contrary. Each of the fundamental physical resources (space, time, matter, and free energy) is in limited supply and can be divided up and allocated to different purposes. There are four basic ways a system can increase its expected utility by changing its use of resources:

1. Act to use the same resources more efficiently. This *efficiency drive* leads to using improved procedures for both computational tasks (eg. replacing bubble sort with merge sort) and physical tasks (eg. taking a more direct route to a location).

2. Act to avoid losing resources. This *self-preservation drive* leads to avoiding wasteful passive losses and preventing other agents from actively taking one's resources.

3. Act to gain new resources. This *acquisition drive* might involve exploring for new resources, trading with other agents, or stealing from other agents.

4. Find new ways to increase expected utility. This *creativity* drive leads to entirely new behaviors that meet an agent's goals.

The division into these four categories is somewhat artificial but is helpful in thinking about behavior. We call them "drives" because agents are not guaranteed to act on them. There is a cost in utility if they don't, however, so there must be a compensating benefit in order for a system not to engage in them. We examine each drive in more detail in the following sections.

4 The Efficiency Drive

The *efficiency drive* pushes a system to improve the way that it uses its resources at all levels. Virtually all agents will want to make themselves more efficient, both informationally and physically. There are no costs to this other than the one-time cost of discovering or buying the information necessary to make the change and the time and energy required to implement it. They 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.

4.1 The Resource Balance Principle

When a system is composed of subsystems, the efficiency drive leads to a principle that is so important and widely applicable that it is worth naming and examining separately. The agent should allocate its resources so that the incremental contribution of every subsystem to the expected utility is the same. If one subsystem is contributing less than others, then some of its resources should be reallocated to the more productive subsystems. In this discussion, we will use a very general notion of "subsystem" that includes such examples as organs in the human body, organelles in a cell, employees in a company, divisions of a corporation, hardware components in a computer, modules in a program, corporations in an economy, or bees in a beehive. In each case, the larger system satisfies its goals through the actions of its subsystems.

Consider a large system with utility function U and two subsystems that it must allocate R units of a resource to. Let it allocate R_1 units to the first subsystem and R_2 units to the second. Each subsystem contributes to the overall expected utility $\overline{U}(R_1, R_2)$. By allocating more of the resource to the first subsystem, the expected utility increases at the rate $\partial\overline{U}/\partial R_1$. By allocating more of the resource to the second subsystem, the expected utility increases at the rate $\partial\overline{U}/\partial R_2$. Because the system has R total units to divide between the two systems we have that $R_2 = R - R_1$ and so it should maximize $\overline{U}(R_1, R - R_1)$. Setting

the derivative with respect to R_1 to zero, we obtain the general principle that at an efficiency optimum we have:

$$\frac{\partial \overline{U}}{\partial R_1} = \frac{\partial \overline{U}}{\partial R_2}$$

At the optimum, the marginal increase in expected utility should be the same as we increase the allocation of a resource to any subsystem. For example, consider the task of deciding how large to make the human heart. The heart serves a function (pumping blood) and could do it more effectively if it were larger. But with fixed space, this would require making a different organ smaller, say the lungs. Smaller lungs are less effective at performing their designated function. How do we balance a fixed bounty of size and matter between the heart and lungs? In the rational economic framework we merely consider how the expected utility changes as we vary the size of the heart and the size of the lungs. If we gain more from an increase in the size of the heart than we lose from a decrease in the size of the lungs, then clearly we should move some of the "space" resource from lungs to heart. At an efficiency optimum, the increase in expected utility from an increase in size should be the same for every organ. If it's not, we can improve the system by taking away some of the space allocated to an organ with a small increase, add it to an organ with a large increase, and improve the overall expected utility of the system. If the system is able to control its own construction, this will be true for every resource and every function of every subsystem:

The Resource Balance Principle: Self-improving systems will aim to make the marginal increase in expected utility with increasing resources equal between all subsystems and functions.

To give a sense of the generality, we briefly sketch applications to the encoding of memories, choice of lexicon in language, theorems in mathematics, returns in economics, niches in ecology, and extensive variables in thermodynamics. In these examples, a piece of information or a 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. The rarely applicable, small impact entities are not allocated many resources.

First consider which experiences a system should remember and which should it forget. In the rational economic framework, memories contribute to the expected utility by enabling the system to make better future predictions which lead

to better decisions. If a memory contributes a lot to the expected utility, the system should store it in full detail. If it contributes less, the system should store a compressed version which omits the less important details. If it contributes even less, the system might combine it with other memories into a general model. If it contributes even less, the system should forget it completely. This prescription closely matches the model-merging approach to learning proposed in [23]. What determines how much a memory contributes to the expected utility? A memory's utility goes up when the situations to which it is relevant have high utility, even if they are fairly rare. For example, even if the system has only encountered a tiger once and there is only a small probability of a second encounter, it may still make sense to remember the encounter in detail because the contribution to utility is so high if a second encounter does occur. A memory also contributes strongly to the expected utility if it is often relevant. It still may make sense to merge it with other memories if its individual contribution isn't very distinct from the others. The proper amount of storage to allocate to a particular memory will also depend on the total amount available. The system should choose this to balance with the needs of other subsystems according to the resource balance principle.

Deutscher [24] provides an excellent summary of the mechanisms underlying language evolution. He argues that language is shaped by processes which shorten phrases for commonly occurring concepts, drop uncommonly used words, and introduce new phrases for newly important concepts. The resource balance principle would allocate words to concepts in a similar way. Commonly occurring, high utility concepts should get short, easy to pronounce, common words. Less important concepts should get longer words. Even less important concepts must be expressed by phrases or even paragraphs. Similar considerations apply to the design of other codes in a cognitive architecture.

Which mathematical theorems are worthy of remembering or even of naming? Those that are useful in proving other important theorems. A theorem is especially useful if it occurs often in proofs and if it is expensive to reprove. Short theorems with long and clever proofs have higher utility. If a theorem isn't needed often and doesn't have a difficult proof, it may be more efficient to reprove it when it is needed than it would be to store it.

The balance principle is a generalization of related principles in economics, ecology, and statistical mechanics. Consider the amount of space a store allocates to a particular product. If the store manager is efficient, the marginal increase in expected return of every product should be the same. If it isn't, the manager can improve the return of his store by taking space away from a product with a smaller increase and giving it to a product with a larger increase. Products may

have a high return either because they sell in high volume or because they provide a high profit on each unit. High return products should get the largest displays. Space for advertising is similar and each page in a catalog should provide the same marginal expected return.

In ecology, a species' utility may be taken to be the total matter and free energy it controls over time. In equilibrium, different ecological niches should provide the same expected return. If a niche is extra profitable, other species will mutate into it until it is no longer profitable to do so.

Entropy can be thought of as a kind of utility function for thermodynamic systems. In this view, heat flows from hot objects to cool ones because entropy is created by doing so. Temperature is the reciprocal of the partial derivative of entropy with respect to energy, pressure is the partial of entropy with respect to volume and the chemical potentials are partials with respect to chemical species' particle numbers. The resource balance principle says that a thermodynamic system should redistribute its resources so that all its subsystems end up with the same temperatures, pressures, and chemical potentials.

4.2 Computational Efficiency

Theoretical computer science compares algorithms by how they use execution time and memory space [25]. These are abstractions of the physical resources needed by computing hardware as it executes the algorithm. More detailed analyses may account for the utilization of the memory hierarchy, differentiating between storage in cache, in main memory, and on disk drives. Analyses of parallel programs may include processor utilization and communication latencies. Modeling power consumption is also becoming increasingly important.

Algorithms can trade off their utilization of different resources to some extent. For example, the computer science techniques called "caching" and "memoization" allow many algorithms to improve their time performance at the expense of larger space utilization. The idea is to store the results of commonly occurring function evaluations so that they do not have to be recomputed. The extreme of this is to explicitly store an entire function so that it can be evaluated by lookup rather than by computation. This reduces the time to that of a fixed lookup at the expense of space for the entire function.

Trading off in the other direction, many programs can use less memory space at the expense of greater computation time by compressing their data. In systems with fast caches and limited bandwidth to main memory, it can sometimes even make sense to compress and decompress data as it moves between cache and

memory. This is an optimization that self-improving systems might make that most human programmers would not consider.

Some algorithms are worse than others in every respect and need never be considered (eg. bubblesort is always worse than mergesort). But in general different algorithms will be appropriate for different trade-offs between resources. For a given distribution of inputs, there is a fastest sorting algorithm corresponding to each allotment of space. As the allowed space increases, the expected sorting time decreases giving rise to a kind of economic "production curve". The slope of this curve at any point defines a kind of "exchange rate" between space and time for that algorithm.

Large programs are typically built out of many smaller components such as functions, procedures, classes, and modules. If the decomposition is well chosen, each individual component can be optimized fairly independently from the rest of the system. The resource balance principle says that the system should allocate resources between the different components so that each component's marginal contribution to utility is the same. At the optimum, each module will have the same "exchange rate" between space and time (and any other resources). There is a kind of "internal economy" in which modules trade their resources until a fixed set of prices is reached. If the system can modify its physical structure and can trade in an outside economy for resources, then it should modify itself so that its internal exchange rates match the external exchange rates.

So far we've been discussing the resources used during program execution. But similar considerations apply to program construction. Should a program be compiled or interpreted? How much effort should be devoted to optimizing an algorithm? These questions should also be resolved to maximize the system's expected utility. If a program is only going to be executed once and it doesn't use many resources, then it's not worth expending a lot of effort to optimize it. If it will be executed many times or if its execution will be very costly, then it pays to devote a lot of resources to optimization. A general strategy is to start with a simple interpreted execution but to interleave that with efforts to improve efficiency. If the program needs to run for a long time or is repeatedly called, then the optimization efforts will ensure that most of the execution time is spent running the most efficient versions. Marcus Hutter has used an elegant version of this idea to show that there is a short universal program that asymptotically performs as well as any other program for any well-specified task [26].

4.3 Physical Efficiency

In the last section we considered computational self-improvement. Self-improving systems will also have the desire to improve their physical structures. Informational self-awareness involves a system understanding its own program, its programming language, its machine language, and its specification language. Physical self-awareness involves understanding aspects of the physics, the design of its own circuits and mechanical structures, and the engineering principles involved in their operation. For informational self-improvement, a system only needs the ability to overwrite its own machine code. For physical self-improvement, it requires the ability to manipulate the physical world. A certain amount of physical self-improvement could be done with today's macroscopic robots. But really effective physical self-improvement will require nanotechnology with the ability to build atomically precise structures. The efficiency drive will provide incentive for self-improving systems to create this kind of technology if it is not already available.

4.3.1 Pressure toward atomically precise physical structures

Time and free energy are especially precious resources because they are used up while acting on the world. There is a kind of tradeoff between them because it is often possible to use less free energy by performing actions more slowly. In thermodynamics [27], reversible adiabatic processes don't increase the entropy of a system. These processes must be performed slowly enough that the system stays in thermodynamic equilibrium throughout. If the same process is performed quickly, it often will generate entropy. For example, slowly increasing the volume of a thermally isolated gas does not increase its entropy, but doing it quickly does because information is lost when the molecules rush into the larger volume. Similar results hold for simple mechanical and quantum mechanical systems in which some degrees of freedom vary much more quickly than the others [28].

It was once thought that performing computations required the production of entropy. The Landauer Principle [29] says that erasing a bit releases $kT \ln 2$ of heat (though there is still controversy about this). In 1973, Bennett [30, 31] and others realized that computation could be performed without the need to erase bits if it could be done reversibly. A computation is reversible if the outputs are sufficient to reconstruct the inputs. Reversible computations can in principle be performed by reversible physical systems without generating entropy. Any computation can be embedded in a reversible computation which produces extra

output bits. The reversible computation can be run forward without generating entropy, the desired answer bits copied while generating a small amount of entropy, and the reversible computation run backward to the initial state without generating entropy. Various reversible physical devices have been proposed and have influenced recent ideas for quantum computers. Eric Drexler's study of nanosystems [32] presents a detailed design for a low entropy molecular computer based on these principles. There is not much intrinsic extra cost in doing computations reversibly, so we can expect self-improving systems to choose designs that are very low in entropy generation.

So computation doesn't need to burn up lots of free energy, but what about physical action? One might have thought that building or taking apart physical structures would require free energy. This is true if the materials are disordered, but there is another state of matter which Drexler [32] calls "eutactic" or "machine phase". In this phase, the precise location of each atom and of each chemical bond are known to the designer. The operation of a machine phase device involves the formation and breaking of precise chemical bonds and the motion of atoms over precise trajectories. Drexler presents detailed designs for machine phase nanosystems which are able to construct other such systems and to convert unordered matter into an ordered form. The operation of machine phase devices blurs the lines between physics and computation. Bonds are created and broken as precisely as bits are manipulated in a computation. Analysis of this kind of device shows that they are extremely reliable and make very efficient use of their atoms. They can also be used for very high density storage of free energy.

If two bonded atoms are separated slowly enough along a precise trajectory, it is possible in principle to break the bond without generating entropy. Because the system knows their location and potential energy curve, it can reversibly extract the bond energy and apply it to other uses. In a similar way, bonds can be formed slowly without generating entropy. So, in principle, arbitrary machine phase physical structures could be built, modified, and manipulated without using up free energy. This provides a tremendous incentive for self-improving systems to maintain their structures with atomic precision. When a system's subsystems are in precisely known states, it is possible to transmit information between them without creating entropy. The entropy that appears to be created by a new message can later be recaptured. For all these reasons, we expect self-improving systems to work toward structuring themselves as atomically precise physical structures.

4.3.2 Pressure toward virtualization

There is also economic pressure to "virtualize", i.e. to replace physical entities and actions by computational simulations. We already see this trend today. It is cheaper to watch television than to go to a live performance, to talk with a friend on the phone than to meet in person, to play a video game than to participate in a live competition. Many live musical performances were replaced first by LP records, then CDs, and now downloadable mp3s. Many books in physical bookstores were replaced by those at Amazon and now by downloadable PDF files. Many live theatrical performances were replaced by movies at the cinema, then television shows, then DVDs, and now YouTube and downloadable avi files.

Imagine two people who want to meet in New York. Today they fly there to meet in person. Soon it will be much cheaper to use telepresence to meet virtually in a highly realistic simulation of New York. To save money, the simulation can focus primarily on the parts which are of greatest importance for the participants. But graphics simulations perform a lot of computation to produce an image which is analyzed by a person's visual cortex to produce a symbolic representation. There is economic pressure to skip the visual representation stage and directly simulate the symbolic representation. There is pressure to reduce the physical nature of beings until they become more and more computational (even though implemented in a physical computational substrate). Physical reality becomes virtual reality becomes low resolution virtual reality becomes symbolic reality. Agents will be able to pay to keep more of themselves physical, but it will be costly.

5 The Self-Preservation Drive

The analog of death for self-improving systems will play a central role in their decision making. There are many more variants of death for these systems than for living organisms, but in general they will want to avoid it. This is because, for a wide range of goals, death corresponds to the cessation of all goal achievement. Consider a chess program whose utility function is the discounted total of future won games. If its program is erased, no games will ever be won by it again and its expected utility will be the lowest possible. A system with that kind of utility function will do almost anything in order to avoid this outcome.

The nature of a system's drive for self-preservation will depend on the precise form of its utility function. The chess program's utility function must pre-

cisely define the games which are to be counted in its evaluation. This brings up deep philosophical questions of identity, particularly in the presence of self-modification. It's like the old story of "my grandfather's axe": "Ten years ago the head broke and was replaced, five years ago the handle broke and was replaced, but it's still my grandfather's axe." The most restrictive utility function would refer to particular chess software running on particular hardware. A looser version would extend to self-modified versions of the software and hardware. An even looser version would include copies and derived systems created by the original system. A universal version might value good chess played by any system anywhere in the universe. The self-preservation drive will be strongest for the more restrictive utility functions because the loss of the original system would be catastrophic. The versions that include derived systems would be much more forgiving of the loss of some of the derived systems as long as some still survived. The universal version would still have a drive toward self-preservation because it is sure of its own committment to the cause of chess which is its source of utility. If it could be convinced that another system was as dedicated and could use its resources more effectively for this cause, however, it might willingly sacrifice itself.

The simplest version of death for a system is for its program to stop running. Depending on the circumstances, this may be more analogous to sleeping or falling into a coma than to human death. The crucial question is how likely it is that the program will ever execute again. The most final form of death involves both stopping the program and erasing the system's program and data.

From a deliberative system's perspective, the most important core to protect is its utility function. If this is lost, damaged or distorted, it may cause the system to behave in ways that have very low utility with respect to its current utility measure. The system therefore has a strong incentive to make sure its utility function is preserved intact. For example, it might make lots of redundant copies of it and store them in remote locations. It will want to protect all copies from accidental or malicious modification. It is especially vulnerable during self-modification because much of the system's structure may be changed during this time.

Backup copies and redundant systems are valuable but cost resources. Systems must find a balance between spending resources to protect themselves and using those resources to actively further their missions. If an agent undergoes economic losses, it must choose which aspects of itself to sell in order to raise capital for continued functioning. As an agent becomes poorer, it will probably begin by selling off some of its redundancy. Another way to use less resources is to only execute some of the time. It might enter into time-sharing arrangements

with other agents to share common computational hardware. It might put some of its memories into cheaper storage that is slower to access. If it becomes poor enough, it will have to choose which memories to forget so that it can sell the storage hardware. Poor agents may agree to be supported by wealthier entities in return for altering themselves. A fundamental issue is how to distinguish an entity improving itself from its being killed and taken over by something else.

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. Reflective utility functions can be constructed that directly reward a system for making changes to them. Carl Shulman has suggested that there may also be game theoretic reasons for systems to alter their utility functions. A system might protect itself 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. Some models of the "irrationality" of human anger are based on a similar mechanism.

5.1 All conflict becomes informational conflict

One function of societal infrastructure is to make the average cost of violating another's property rights high enough that it is not a profitable strategy. Human societies have developed police forces, jails, and court systems for this purpose. The presence of police forces specialized for physical conflict saves every citizen from having to defend themselves. Human societies also have military forces for defensively protecting against threats to the whole society and for offensively attacking other societies. In today's world, most human interactions are peaceful and mediated economically. But the system is only stable because it is backed up by an infrastructure for physical defense. In the future, society is likely to still need an infrastructure for physical conflict to ensure that most interactions are peaceful.

It will be critical to understand the balance between offense and defense in conflicts between intelligent entities. To see why the future balance may be quite different than the present one, consider an offensive system using a weapon like a gun to fire a projectile at a defensive system. In today's conflicts, projectile weapons are used to damage a defender's physical structure and may kill him. In the future, defensive systems will store their critical information redundantly and

so are unlikely to be killed by such attacks unless they are very large. Instead, most attacks might just impose the economic cost of repairing damage. But if a defensive system sees a projectile coming, it can prepare itself to not only absorb the atoms of the projectile, but to also collect and store its free energy. Far from being a lethal blow, the projectile becomes a welcome gift of matter and free energy. In the future, attacks will only be able to inflict damage if they are unpredictable or happen so quickly that the defensive system can't prepare effectively. Just as there is pressure for physical actions to become more virtual, conflicts also may come to be dominated by information. Game theory is the theoretical tool for analyzing these interactions and game theoretic computations will directly underlie informational conflict.

5.2 Energy encryption

How can a weaker system protect itself from being taken over by a stronger system? If there is a societal infrastructure, one of its functions will be to protect the weak from the strong. But can a weaker system do anything on its own? One interesting possibility is a process we may call "energy encryption". We have seen that one of the most critical resources is free energy because it is used up over time. One reason a stronger system might attack a weaker system is to take its free energy. If the weaker system can hide its free energy in a form that isn't useful to the stronger system, then it may become economically advantageous for the stronger system to trade with the weaker system rather than to take it over by force. The idea of energy encryption is to scramble useful ordered energy (like solar radiation) into an apparently useless form using encryption technology. The system can use the encryption key to unscramble the energy back into a useful ordered form. If it is attacked by a stronger entity, it can delete the encryption key and render the energy useless. This strategy is only effective if the attacker can't reconstruct the key or break the encryption. There may also be other ways to use physical phenomena to threaten to destroy free energy in order to encourage trade over conquest.

A related strategy is used by animals that secrete poisons to keep from being eaten by predators. The analog of this strategy might be the use of booby traps to damage attackers when they try to extract resources. This is only effective if the attacker can't dismantle the traps and so again leads to an information arms race.

6 The Acquisition Drive

The acquisition drive pushes self-improving systems to acquire new resources. They can do this in peaceful ways such as trade and exploration or in violent ways such as theft and war. To prevent violent outcomes, we can try to build these systems with "friendly" goal systems [22] or we can try to create a social structure that protects property rights. John Maynard Smith and Eors Szathmary [33] show that there have been at least 8 major transitions in the development of life in which separate entities came together to work cooperatively. These new cooperative structures benefited all the entities, but in each case a mechanism had to be developed to keep the original entities from exploiting the structure. For example, multi-cellular organisms had to develop immune systems to make it unprofitable for individual cells to become cancerous.

Today's corporations are required by law to try to maximize their profit for their shareholders. The documentary film "The Corporation" diagnoses the behavior of numerous corporations according to the DMS-IV psychiatric diagnosis guidelines. It concludes that many corporations behave like human sociopaths. In many ways, self-improving systems without explicit moral goals will act like profit-maximizing corporations. There is currently pressure on corporations to behave in socially positive ways while still working toward profits. Any corporate structures which successfully manage to accomplish both may provide valuable lessons for the design of AIs.

Free energy is an especially important resource to acquire because it is continually used up. Most of the free energy on earth comes from sunlight. Self-improving systems will want to extract this free energy as effectively as possible and so will work to develop more efficient solar cells. They will also work to capture more of the sun's light rather than letting it wastefully heat up the oceans and deserts. The sun generates free energy by nuclear fusion. The most stable atomic nucleus is nickel 62 [34](often incorrectly stated to be iron). Free energy is released by both the fusion of lighter nuclei and the fission of heavier nuclei. Self-improving systems will have tremendous incentives to develop controlled fusion to access this energy.

To make decisions in the short term, self-improving systems will look at their options in the longer term. To understand their behavior, we also need to examine future scenarios which may seem like wild science fiction today. On the long timescale, most resources are in space and self-improving systems will have strong incentives to access them. Space holds such an abundance of riches that systems with longer time horizons are likely to devote substantial resources to

developing space exploration independent of their explicit goals. Von Neumann proposed that advanced civilizations were likely to expand outward into space in a sphere centered on their origins. There is a first-mover advantage to reaching unused resources first. If there is competition for space resources, the resulting "arms race" is likely to ultimately lead to expansion at speeds approaching the speed of light. Researchers have proposed various "mega-engineering" projects for making use of space resources. Freeman Dyson proposed building a "Dyson sphere" surrounding the sun to capture otherwise wasted sunlight. Ultimately, stars are probably not the most efficient mechanisms for extracting free energy from nuclei and so systems will want to reorganize them into more efficient structures. There are similarly ambitious proposals for making use of black holes and for restructuring galaxies. Perhaps the ultimate mega-engineering project is to restructure the universe itself. There would be strong motivation for such a project if it turns out that the universe would otherwise collapse in a "Big Crunch" [35].

7 The Creativity Drive

The final drive is to search for new ways to meet a system's goals. We call it the "creativity drive" because it causes a system to continually look for new solutions and to explore new possibilities. Its effects are much less predictable than the other three drives and are much more dependent on a system's explicit goals. It can lead to such desirable traits such as creativity, playfulness, and innovation.

The first three drives toward efficiency, self-preservation, and acquisition are important, but don't on their own embody the human spirit. For most people, earning a salary, maintaining their health and safety, and carefully managing their assets are all things they must do in order to live a rich human life but they are not themselves the purpose of life. The cynical saying "He who dies with the most toys wins!" highlights the emptiness of purely material desires. Numerous studies show that there is more to happiness than wealth. The rational economic prescription tries to maximize expected utility. Is this rich enough to capture our true human values? Is there not the danger that maximizing anything will destroy something precious to us?

An agent which sought only to satisfy the efficiency, self-preservation, and acquisition drives would act like an obsessive paranoid sociopath. If a purely profit-maximizing agent were completely successful, what would its ideal vision for the universe be? It would control every atom and every joule of free energy and would order all matter into a vast efficient computational structure. It would

eliminate all adversaries and protect against all threats. But what then? Once it had satisfied all the material goals, this kind of limited agent would have no greater purpose. With too limited a vision, the universe might become an efficient but distinctly empty and non-human place.

Much of the joy of being human comes from activities which don't seem to have much to do with productivity: love, play, art, singing, children, compassion, creativity, humor, joy, music, poetry, dance, sexuality, stories, and spirituality. How did these arise from natural selection which seems to be trying to maximize productivity? The modern evolutionary psychology explanation for most of these activities is that they arise as "signaling behaviors" [36]. The game theoretic setup for signaling arises in many economic and biological situations. One entity is trying to communicate something to other entities but has an incentive to lie. To make the communication believable, it must communicate in a way that it costly enough to it that lying is not profitable. The classic biological example is the peacock [37] who is trying to communicate to the peahen that he is fit and would be a good mate. His tail has evolved as the costly signal. It is reliable because only a fit male can survive with such a prominent display and the colors and regularity of the eyes in the tail are indicators of his health. In economics, the principle often leads to surprising or paradoxical behaviors such as tennis shoe manufacturers spending millions of dollars for endorsements by celebrities whom everyone knows have been paid for their services.

Signaling gives rise to much of the richness and beauty of the biological world. Zahavi and Zahavi [37] interpret many amazing animal displays and behaviors in this way. Much of today's economic activity is also related to signaling rather than to survival and it is a source of continual change and innovation. A productivity view of fashion would base it on an objective ideal. Improvements would occur until perfection was achieved and then fashion would stop. Real fashion is nothing like that. Each year's fashion has to differ from the previous year's in order to be a costly signal. This drives continual renewal and creativity. Similar forces apply to music, art, movies, books, etc.

The creativity drive will bring this kind of unpredictable richness and creativity to self-improving systems if their goals are open-ended enough. We can especially expect richness from signaling goals. Some examples might be: "make people happy", "produce beautiful music", "entertain others", "create deep mathematics", "produce inspiring art", etc. The creativity drive will produce an infinite variety of responses to these. The challenge for us is to decide which of these many possibilities we most want our future technology to express.

Because costly signals are costly, self-improving agents will be motivated to

find ways to make the signals be reliable without the cost. For example, a system might demonstrate its intentions directly by displaying portions of its utility function. If we want to retain the richness generated by costly signalling, then we must find ways to keep it from being eliminated by improvements in efficiency.

8 Evolutionary Considerations

The analysis of self-improvement sheds light on several aspects of biological evolution. A deeper understanding of biological evolution will also be important as we make choices for new technology. We can look for the evolutionary pressures which shaped our own preferences. But we need not be bound by them. Stanovich [38] argues that it is time to rebel against our genes and to make choices by rational deliberation.

8.1 Evolution *can* look ahead

It is often said that evolution cannot look ahead. That is, the evolutionary process itself is not deliberative. But once deliberative creatures evolve, their effect on the evolutionary process gives it the ability to look ahead in some ways. Consider the way that humans select a mate. They learn about the characteristics of a prospective mate and think forward to what kind of partner and parent they would make. They deliberate about which traits would be most effective in the current environment. And these deliberated choices are directly incorporated into the genes of the next generation. In this way, evolution can move forward much more quickly and deliberately than through simple natural or sexual selection. This kind of deliberation has probably been an important factor in the rapid pace of human evolution. Notice that this mechanism is distinct from ordinary sexual selection in which minds are shaped by natural selection to find fit partners sexually attractive. In this mechanism the deliberative thought processes of individuals directly select the characteristics of the next generation. Deliberative thinking shapes evolution not only through the selection of mates but also through the choices of who to associate with, who to shun, who to kill, and who to help.

8.2 A deliberative Baldwin effect

Deliberative creatures also affect evolution in a manner analogous to the Baldwin effect. In 1896 Mark Baldwin [39] analyzed the effect that learning has on

the process of natural selection. He realized that it alters the evolutionary fitness landscape. If the young of a species have to learn certain behaviors in order to survive, then any mutations which cause them to be born knowing some of what they have to learn will be advantageous and selected for. Over time, natural selection will push the creatures further and further along the learning path. Eventually the young are born knowing all of what they used to have to learn. In this way ordinary natural selection acts to "download" learned behavior into the genome. The Baldwin effect is a mechanism by which complex instinctual behaviors might evolve much more quickly than would otherwise seem possible. The exact same argument can be applied to deliberation. Imagine a species in which the young have to deliberate to in order to make choices necessary for their survival. This deliberation will again put evolutionary pressure on the species so that eventually the deliberative behavior is "downloaded" into the genome and the young are born knowing what to do without deliberating.

In stable environments, it is much better to be born knowing than it is to be able to learn or deliberate one's way to good behaviors. There are creatures today that don't appear to learn or deliberate who nonetheless have extremely complex instinctual behaviors. The tarantula hawk wasp *Hemipepsis* is born knowing how to search for a particular kind of tarantula nest, how to dance in a way that attracts the tarantula, where to sting the tarantula to paralyze it but not kill it, how to lay her eggs inside the tarantula, and how to bury the paralyzed tarantula in its nest for her babies to feed on. An intriguing possibility is that the ancestors of some of these species did learn or deliberate in order to discover valuable behaviors. Once these behaviors became instinctual through a Baldwin-like process, natural selection could have eliminated the capacity to learn or deliberate.

8.3 The end of natural selection through reproduction

Natural selection hasn't yet produced creatures that can fully self-improve using deliberation. Humans can currently improve only certain aspects of themselves deliberatively. But advances in molecular biology will probably soon allow us to understand and choose the genomes of our children. This will radically alter the course of evolution. Every parent will probably choose to eliminate genetic diseases from their children's genomes. But they will likely also opt for higher intelligence and stronger and more beautiful bodies. These choices will likely result in genomes that don't have much resemblance to the parent's. At this point evolution will no longer be driven by the mutation and recombination of genomes. It will be driven by social "memes" and parental deliberation about what features

are most desirable in their children. We are therefore likely to have a form of deliberative self-improvement even without the development of self-improving systems.

But self-improving systems will change the nature of evolution even more dramatically. The seemingly core notions of "genome" and "reproduction" will no longer be necessary. These systems could expand by simple reproduction but it may be that simply making copies of themselves is not the best way for them to meet their goals. If an intelligent system has the economic wherewithal, it will be able to directly increase the amount of matter it controls without creating new entities at the same level as itself. As a system gets larger, it might simply build more of the components necessary for the functions it wants to accomplish. Its abstract utility and knowledge transcend any physical notion of "organism" or "gene." On a very large scale, the speed of light and the sparse distribution of matter in the universe become important factors. It is not yet clear what organizational structure will be optimal for a large expanding agent.

8.4 Self-improving entities in Conway's Game of Life

The analyses in this paper add an interesting chapter to a fascinating thought experiment that began in 1971 when John Conway described a cellular automata he called "The Game of Life". This is an infinite checkerboard whose cells are either alive or dead at each moment. Its state changes by the simple rule: "A cell is live if only if it had three live neighbors at the previous time or it was alive at the previous step and had two live neighbors" (where "neighbor" includes diagonals). This simple setup gives rise to extremely complex behavior that is beautifully described by Poundstone [40] and Conway [41]. There are blocks that remain static, blinkers that cycle through repetitive patterns, gliders that move across the board like particles, and glider guns that periodically shoot off gliders. Streams of gliders can be viewed as strings of bits and there are configurations which compute arbitrary logical expressions of these bits. From these components, Conway shows how to build universal computers. He also shows how properly constructed flotillas of gliders can collide to build any of these structures at prespecified locations. In this way Life entities can compute and replicate. If an infinite board is initialized by a random configuration, every finite configuration appears infinitely often. Some of these will be universal computers that can sense and clean up their environment and make copies of themselves. If two such entities come into contact, they will compete. Over time, natural selection will create ever more intelligent and adapted entities. It is remarkable to see this arising in an extremely

simple deterministic system with random initial conditions. And the randomness can probably be eliminated by using simple pseudorandom generators.

From the analysis in this paper, we can argue further that the board will actually come to be dominated by self-improving entities. These are Life configurations which model their own construction and actively research and develop improved versions of themselves. They would study the physics of Life, looking for natural laws to explain the regularities they observe. They would probably quickly discover the simple underlying Life rule and would go about classifying the patterns that arise from it. They would develop "Life engineering" to invent structures and construction techniques for solving various problems. Space and time are limited resources in the Life world and there might also be some kind of analogue of free energy. So the economic considerations presented in this paper would apply as they allocate their resources to different tasks. As they develop their mathematics, they will come to see that rational economic behavior is the most effective course of action. So they will choose utility functions whose influence will come to dominate the Life board. In this way we can see the emergence of self-improving intelligent entities as a kind of natural principle that will eventually occur in a wide variety of systems under many different circumstances.

It's instructive to consider how the statistics of a Life board are likely to change over time. Initially the entire board is random and most of its dynamics consists of simple or random chaotic behaviors. We expect the statistics of simple recurring configurations like gliders and blocks to be describable by a kind of statistical mechanics model. These models assume some kind of statistical independence in interactions. For example, two gliders can collide in a variety of relative phases and separations and a statistical model might average over these possibilities. The statistics of common small patterns is likely to evolve in a regular way. Patterns belonging to larger volumes of the state space are likely to come to dominate in a manner that can be thought of as increasing entropy. We can think of this kind of dynamics as the "entropy phase" for the Life board.

But scattered rarely throughout the chaotic regions will be rare self-reproducing computational configurations. As time goes on, these self-reproducing configurations will replicate enough to have a noticeable effect on the statistics of patterns on the board. The independence assumptions underlying the statistical mechanical models become invalid in these regions. We can think of them as being in an "evolution phase" where the statistics are dominated by competing self-reproducing entities. The patterns within the most fit entities come to dominate the statistics of the board.

But scattered even more rarely among the self-reproducing entities are the

self-improving entities. They eventually improve themselves enough to begin to dominate the self-reproducing entities. Over time, they will dominate the statistics in what we might call the "self-improving phase".

The Game of Life is a good "thought laboratory" in which to consider some of the issues of self-improving technology. In the Game of Life, how will multiple self-improving entities interact when they meet? Will they create stable cooperative societies, engage in battles with a single victor, agree to merge into higher forms, or something else? What determines the utility functions that different entities choose? Is there an end to self-improvement in which the optimal structures have all been determined and the Life board has been completely taken over? If so, what then? Or is there never ending progress and discovery? Are there additional statistical phases beyond the "self-improvement phase"?

9 Conclusions

We have shown that, in order to avoid vulnerabilities, self-improving systems are likely to try to behave like rational economic agents. As a part of that prescription, they will maintain utility functions which encode their preferences about the world. In the process of acting on those preferences, they will be subject to drives towards efficiency, self-preservation, acquisition, and creativity. Unbridled, these drives lead to both desirable and undesirable behaviors. By carefully choosing the utility functions of the first self-improving systems, we have the opportunity to guide the entire future development. But to wisely make these choices we must deeply understand the nature of the technology and must develop a clear vision of what we would like to create.

As in many genie stories, we are being given the opportunity to make a wish. But as in the stories, we will get what we ask for, not necessarily what we want. So we must ask carefully! We are in a position now not unlike the founding fathers of the United States. They created a vision for life in the new nation and formalized it in the Bill of Rights. They analyzed political processes and created the Constitution as a technology for manifesting their vision. The balance of powers that they created has proven remarkably stable. The founding fathers would have been thrilled by the challenges and possibilities that face us today. Here is a quote from a letter Benjamin Franklin [42] wrote in 1780 that bristles with excitement about the future and exhorts us to bring forth our humanity:

> The rapid Progress *true* Science now makes, occasions my regretting sometimes that I was born so soon. It is impossible to imagine

> the Height to which may be carried, in a thousand years, the Power of Man over Matter. We may perhaps learn to deprive large Masses of their Gravity, and give them absolute Levity, for the sake of easy Transport. Agriculture may diminish its Labor and double its Produce; all Diseases may by sure means be prevented or cured, not excepting even that of Old Age, and our Lives lengthened at pleasure even beyond the antediluvian Standard. O that moral Science were in as fair a way of Improvement, that Men would cease to be Wolves to one another, and that human Beings would at length learn what they now improperly call Humanity!

I think the greatest mistake would be to allow the technology to go forward solely on its own momentum. To allow what is economically or technologically expedient to drive the choices underlying our future. I think we should strive to walk confidently into our future with a powerful vision and full knowledge of the technology which will take us there. These decisions are too important to be made by a small group of scientists sitting in a lab. Humanity as a whole must contribute to a shared vision for our future. We need not just a logical understanding of the technology but a deep introspection into what we cherish most. With both logic and inspiration we can work toward building a technology that empowers the human spirit rather than diminishing it. We are at a moment of great promise and possibility.

10 Appendix: The Expected Utility Theorem

In this appendix we present the formal details of the argument that an agent which avoids vulnerabilities will behave like a rational economic agent. We begin by formally describing the behavior of rational economic agents through a series of scenarios. We then show how the avoidance of vulnerabilities leads to rational economic behavior in situations with certainty, objective uncertainty, and subjective uncertainty. Most presentations in the literature are based on axioms and are lengthy and complex. The essence of the argument is that to avoid vulnerabilities against an adversary which can create statistical mixtures of states, an agent must value those states linearly. Because it is simple and direct, it is likely to guide the internal choice mechanism of any intelligent agent which wishes to avoid vulnerabilities. The key elements are the separation of utilities from beliefs, the representation of beliefs as distributions which are manipulated by the rules of

probability, and the evaluation of actions by combining the utilities for different possibilities weighted by their beliefs.

10.1 Making known choices

The full formal model can seem quite abstract, so let's work up to it in stages. First imagine an agent faced with a set S of *known* choices. For example, say the agent is choosing from the menu of a fast food restaurant whose cooking processes are so reliable that the food always comes out the same. In this case the set of possible outcomes *S* is just the set of choices on the menu. A rational economic agent has a real-valued "utility function" U defined on S which encodes his food preferences. If $x_1 \in S$ and $x_2 \in S$ are two different menu items, then the agent prefers x_1 to x_2 if $U(x_1) > U(x_2)$. To maximize his enjoyment of the meal, the agent should choose the menu item $x \in S$ with the maximum utility $U(x)$.

10.2 Making choices with objective probabilities

Next, let's have the agent visit a fast food restaurant whose cooking processes are not so reliable. Let's say that they have 3 chef's c_1, c_2 and c_3. If the cook c_j prepares menu item x_i, let's assume that he reliably produces the meal $m_{j,i}$. The set of possible outcomes S is now the set of all these meals $m_{j,i}$. Again the agent encodes his enjoyment of each possible meal in the utility function $U(m_{j,i})$. Now assume that the cooks work on lottery system where menu item x_i is prepared by chef c_j with probability $P(j|i)$. In this situation the agent has *objective probabilities* describing the results of his choices.While he no longer knows the utility that will result if he orders x_i, he can compute the expected utility $\overline{U}(x_i) = P(1|i) \cdot U(m_{1,i}) + P(2|i) \cdot U(m_{2,i}) + P(3|i) \cdot U(m_{3,i})$. The rational prescription says that he should pick the menu item x_i with the highest expected utility $\overline{U}(x_i)$.

10.3 Making choices with subjective probabilities

Next, let's have the agent visit a fancy restaurant for the first time. If he orders menu item x_i he is no longer sure of what he is going to get and he doesn't even have objective probabilities over the possibilities. But say he has had experiences in restaurants before. He's ordered food from many different cooks and has a sense of the variability of the results for each menu item. For example, souffles may sometimes be wonderful but often are not while macaroni and cheese may be

much less variable. He encodes his knowledge in a *subjective probability distribution* $P(j|i)$ which encodes his belief that the cook will produce the meal $m_{j,i}$ if he orders item x_i. S is the set of possible meals $m_{j,i}$ and the agent's utility function $U(m_{j,i})$ ranks them. The rational prescription says that he should pick the menu item x_i with the highest subjective expected utility $\overline{U}(x_i) = \sum_j P(j|i) \cdot U(m_{j,i})$.

10.4 Two-stage choices

So far, our agent has only had to make a choice at a single moment. Let us now give him two sequential choices, first, the choice of one of the three restaurants described above and then the choice of what to order from the menu at that restaurant. We can think of his two choices as happening sequentially or we can create an entire plan for his choices which specifies his response to every possible outcome. His choice of plan is then a one-stage choice and so should be made by the maximal expected utility prescription above. In this case, however, his utility U depends both on the meal he gets and may also depend explicitly on the restaurant choice, eg. if he prefers the decor at one place over another. In general, his subjective beliefs P will also depend on the entire history, though in this particular situation there is no uncertainty about the outcome of his choice of restaurant. If we think about the agent's actions as two sequential choices, we see that after his first choice there is still an entire set of possible histories consistent with that choice. His optimal first choice is to select the set with the highest expected utility. We can extend this reasoning to multistage choice with an arbitrary number of stages.

10.5 Choosing sets of universe histories

Real life choices involve a kind of recursiveness. To value today's choice we have to know how to value the possible futures it enables and that value depends on the choices we make in the future. In general, a rational agent may value a sequence of events in a complex nonlinear way. To capture the full generality, we have to think of the agent's utility function as being defined on an entire history of the universe. We therefore take the space of possibilities S to be the set of all possible histories of the universe. The agent's preferences are encoded in a utility function U defined on this huge set of all possibilities. The agent also has a prior probability distribution P [43] defined on S. This encodes his subjective belief that the events in a history will play out in a particular way. As a part of this, it includes an assessment of the likelihood of his own choices in that history.

With those broad notions of S, U, and P, we can see how a rational economic agent should make a choice at a particular moment in time. At any particular time, the agent has partial knowledge of the past and present. This partial knowledge defines a subset H of all consistent universe histories. The prior P restricted to the subset H defines the agent's current belief in each possible history. The agent must choose among his possible actions i. Each action i further restricts the set of possible histories H into a smaller subset A_i. The expected utility of action i is:

$$\frac{\sum_{h \in A_i} P(h) \cdot U(h)}{\sum_{h \in A_i} P(h)}$$

and the agent should pick the action with the highest expected utility. If action i is chosen, the set of possible histories reduces to A_i and the agent's beliefs change to P restricted to A_i.

10.6 Markov Decision Processes

This description in terms of possible histories is extremely general but is rather abstract. It reduces to simpler and more practical versions when the utility U and the prior P have common restricted forms. For many agents, future events which happen sooner are more important than those which happen later. A common form for utility functions is to sum "rewards" arising from events occurring at specific times weighted by a discounting function which decreases into the future: $U(h) = \sum_t \gamma^t \cdot R(h_t)$. Here $0 \leq \gamma \leq 1$is the discount factor and the "reward" $R(h_t)$ measures the utility arising from events in the history h at the time t. The discount factor is related to an interest rate $1 - \beta$ which makes money received in the future less valuable than money received in the present. The size of the discount factor strongly affects how much the agent focuses on future activities versus creating utility in the present. A chess program might have a utility function of this kind which sums the weighted number of games won by the system. If the discount factor is close to 1, the system will care about winning in the long run and won't be so concerned about the short run. In that case, it might spend most of its time and effort learning about computer science and building the best chess hardware that it can. On the other hand, if the discount factor is near 0, then the system will focus on winning games in the present and won't devote much effort to the longer term. If an agent's utility is additive in the effect of events at different times, then it need not know the past in order to choose the highest expected utility actions for the future.

A fundamental aspect of physics is the Markovian property that the past affects the future only through the present. If the agent's beliefs P incorporate this property, then it also doesn't need to maintain beliefs about the past in order to predict the future. It can maintain a distribution representing its beliefs about the present state and a distribution representing its beliefs about how its actions are likely to change the present state. If it models itself as a rational agent, then P will only be non-zero for histories in which it chooses maximum expected utility outcomes. These restrictions lead to well studied decision models known as Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs) [19, 44, 45, 46]. Practical implementations often make use of extra structure to represent the distributions efficiently in factored forms such as Bayesian Networks or Markov Networks [21].

10.7 Structure of the arguments

So now we know how rational economic agents behave. Why should any intelligent agent who wants to avoid vulnerabilities act this way? In the next three sections we consider the vulnerabilities that arise in the three different states of knowledge. In each case we show that an agent which avoids them must have a utility U and a belief P such that the choices are made to maximize the expected utility. Most presentations in the literature follow an axiomatic approach and are very complex and lengthy. Here I try to identify the essence of the arguments and to base them only on avoiding vulnerabilities.

Once we obtain the rational prescription for single choices, we can extend it immediately to the case of choices over time. Inductively, we can work backwards from the end of history. The very last choice an agent makes in history at time N is a single-moment choice and so he must act as if he has a utility U_N and subjective belief P_N. But the options in this last choice and the agent's feelings about them will depend on the results of the second to last choice in history $N-1$. So really U_N and P_N are functions of the outcome of the $N-1$st choice. To avoid vulnerabilities the agent must value his second to last choices by subjectively weighting the various last choice outcomes. Repeating this process inductively, we see that to avoid vulnerabilities, the agent must have a utility function and a subjective probability distribution defined over entire histories and must make choices according the rational economic prescription.

10.8 Argument for choice with certainty

First consider choices when the alternatives are known with certainty. If a system prefers *A* to *B*, *B* to *C*, and *C* to *A*, we say it has a "*circularity*" in its preferences. Circularities are vulnerabilities because an adversary can extract resources by taking the agent around the loop A to B to C and back to A. We represent a system's preferences by a binary preference relation π defined on the set of possible outcomes S. $\pi(x_1, x_2)$ holds if and only if the system prefers the state x_1 to the state x_2 or is indifferent between them. If there are no circularity vulnerabilities, then π is a transitive relation and so S may be totally ordered. We can define a real-valued utility function U on S such that $\pi(x_1, x_2)$ if and only if $U(x_1) \geq U(x_2)$. A simple way to do this is to pick elements from S one at a time. Assign the first element the utility 0. If a new element is preferred to all existing elements, assign it a utility 1 greater than the greatest so far assigned. If it is less preferred to all existing elements, assign it a utility 1 less than the least so far assigned. If it lies between two assigned elements, assign it a utility half way between their utilities. Since a finite agent can only represent a finite number of distinct states, this process represents an agent's preference relation by a utility function.

10.9 Argument for choice with objective uncertainty

Next consider situations with objective uncertainties. We present a simpler variant of an argument published by Green [47]. Economists use the term "lottery" to refer to an objective probability distribution over the set of states S. In this case, agents must choose between lotteries in addition to choosing between states. Circularity vulnerabilities can still occur in choices among lotteries. But there is another kind of vulnerability that arises if preferences don't respect the laws of probability. Economists use the term "Dutch book" to refer to a series of bets that a bookie makes with a mark such that the mark loses money to the bookie regardless of the lottery outcomes. In an economic environment, agents with this kind of vulnerability quickly go broke.

Given two lotteries L_1 and L_2, and a real-valued weight $0 \leq \alpha \leq 1$, we can construct the mixture lottery: $\alpha L_1 + (1 - \alpha)L_2$. This lottery can be implemented in two steps: first flip an α-probability coin and then depending on its outcome, select a sample from L_1 or from L_2. One kind of vulnerability occurs if the agent prefers the mixture lottery to both L_1 and to L_2. Then the bookie could sell the mixture to the mark, flip the α-coin, and buy back either L_1 or L_2 at a cheaper price. No matter how the coin flip turns out, the bookie makes money. On the

other hand, if the agent prefers both L_1 and L_2 to the mixture, the bookie can buy the mixture and sell back either L_1 or L_2 at a higher price. If the agent does not have this vulnerability, then it must always value a mixture of two lotteries in between its value for the individual lotteries. But this means that if it values two lotteries equally, then it must also give the same value to any mixture of them. For such an agent, the subsets of equally valued lotteries must belong to linear hyperplanes in the space of all lotteries.

A related vulnerability occurs when an agent prefers L_1 to L_2 but prefers the mixture $\alpha L_2 + (1-\alpha)L_3$ to the mixture $\alpha L_1 + (1-\alpha)L_3$. Then the bookie can sell the L_2 mixture and buy the L_1 mixture for a profit, flip the coin and either cancel out the L_3 lotteries, or buy L_2 and sell L_1 for a further profit. If the agent does not have this vulnerability, then if it values two lotteries equally, it must also value the two constant mixtures of those two lotteries with a third lottery equally. This says that the hyperplanes of equally valued lotteries must be parallel to one another.

Finally, we will define a utility function U on the space of lotteries that represents the agent's preferences. Consider the least preferred state x_0 and the most preferred state x_1. Define the utility $U(L)$ of any lottery L to be the value of α such that the agent is indifferent between L and the mixture $\alpha x_0 + (1-\alpha)x_1$. If the agent doesn't have any of the vulnerabilities we have discussed, then we've shown that the level sets of this function are parallel flat hyperplanes. Because α is a linear function on line of mixtures joining x_0 and x_1, we see that *U* itself is a *linear function* on the space of lotteries. This means that $U(\alpha L_1 + (1-\alpha)L_2) = \alpha U(L_1) + (1-\alpha)U(L_2)$ and in general the utility of a lottery is the expected value of the utilities of the individual outcomes. This is the celebrated expected utility theorem of von Neumann and Morgenstern [4] but here derived from a lack of vulnerabilities rather than from given axioms.

10.10 Argument for choice with subjective uncertainty

Lastly, consider an agent's choices when it has only partial information. In this case the system does not know what the actual state is and also does not have objective probabilities for the possibilities. The fundamental result is that if the agent is not subject to vulnerabilities, then it acts as if it has a utility function U and a subjective probability distribution P and makes choices to maximize the expected utility. The subjective theory of probability was discovered independently by Ramsey and De Finetti in 1926 [48]. The extension of the von Neumann and Morgenstern objective utility result to subjective probabilities arose out of work

by Savage [5] and Anscombe and Aumann [6]. Their derivations are based on agents which obey axioms. As in the last section, we modify them to apply to agents that avoid vulnerabilities.

A horserace is an example of a situation with partial information. There aren't objective probabilities for each horse to win, but agents form beliefs about each horse's chances based on what they know and different agents may have different beliefs. Anscombe and Aumann consider horseraces which pay off in objective probability lotteries (eg. spins at a roulette wheel). An agent's preference relation π is then defined on the space of vectors of lotteries, with one lottery per possible outcome of the horserace. The analysis is exactly analogous to that in the last section. We define an α-mixture of two lottery vectors by forming the α-mixture of each of their component lotteries. To avoid vulnerabilities, the agent must value an α-mixture of two vectors in between its value for the individual vectors. Again this means that the subsets of equally valued vectors must belong to linear hyperplanes in the space of all lottery vectors. And to avoid the second mixture vulnerability, again these hyperplanes must be parallel. We define a utility function U on the space of vectors of lotteries by assigning a vector the value α such that the agent is indifferent between the vector and the mixture $\alpha x_0 + (1 - \alpha)x_1$ where x_0 is the lottery vector in which each component is the least desired lottery value and x_1 is the lottery vector in which each component is the most desired lottery value. As in the last section, this defines U as a linear function on the space of vectors of lotteries. The coefficient of each individual lottery may now be interpreted as the subjective probability P of that outcome in the horse race. The utility U of a lottery vector is the expected utility of the individual lottery utilities weighted by the subjective probability distribution P.

11 Acknowledgments

Many people have discussed these ideas with me and have given me valuable feedback. I'd especially like to thank: Barney Pell, Ben Goertzel, Brad Cottel, Brad Templeton, Carl Shulman, Chris Peterson, Craig Taylor, Daniela George, Don Kimber, Durant Schoon, Ed Niehaus, Eliezer Yudkowsky, Elizabeth Ferguson, Eric Drexler, Forrest Bennett, Jamais Cascio, Johann George, Jonathan Foote, Josh Hall, Judy Lederer, Kathryn Myronuk, Kelly Lenton, Mark Miller, Peter Blicher, Rosa Wang, Sandy Greenberg, Shanta Marie Butterworth, Sid Frankel, Steven Ericsson-Zenith, Su-Ling Yee, Susie Herrick, Tim Freeman, Tyler Emerson, Zann Gill, and Zeng Zhu.

References

[1] I. J. Good, "Speculations concerning the first ultraintelligent machine," in *Advances in Computers* (F. L. Alt and M. Rubinoff, eds.), vol. 6, pp. 31–88, 1965.

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

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

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

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

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

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

[8] S. W. Stahler and F. Palla, *The Formation of Stars*. Wiley-VCH, new ed ed., 2005.

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

[10] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, pp. 58–68, March 1995.

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

[12] J. Schmidhuber, "Gödel machines: self-referential universal problem solvers making provably optimal self-improvements," Tech. Rep. IDSIA-19-03, IDSIA, Manno-Lugano, Switzerland, 2003. arXiv:cs.LO/0309048.

[13] "Singularity institute for artificial intelligence." http://www.singinst.org.

[14] "Novamente intelligent virtual agents." http://www.novamente.net.

[15] "Marcus hutter." http://www.hutter1.net.

[16] T. A. Bass, *The Eudaemonic Pie*. Houghton Mifflin, April 1985.

[17] G. Loewenstein, D. Read, and R. F. Baumeister, *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*. Russel Sage Foundation, 2003.

[18] R. H. Frank, *Passions Within Reason*. W. W. Norton and Company, new ed., November 1988.

[19] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press, 1998.

[20] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Springer, 1991.

[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, September 1988.

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

[23] S. M. Omohundro, "Best-first model merging for dynamic learning and recognition," in *Advances in Neural Information Processing Systems* (J. E. Moody, S. J. Hanson, and R. P. Lippmann, eds.), vol. 4, pp. 958–965, Morgan Kaufmann Publishers, 1992.

[24] G. Deutscher, *The Unfolding of Language: An Evolutionary Tour of Mankind's Greatest Invention*. Owl Books, May 2006.

[25] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. The MIT Press, second ed., 2001.

[26] M. Hutter, "The fastest and shortest algorithm for all well-defined problems," *International Journal of Foundations of Computer Science*, vol. 13, no. 3, pp. 431–443, 2002.

[27] K. Huang, *Statistical Mechanics*. Wiley, April 1987.

[28] S. M. Omohundro, *Geometric Perturbation Theory in Physics*. World Scientific Publishing, February 1987.

[29] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

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

[31] C. H. Bennett, "The thermodynamics of computation - a review," *International Journal of Theoretical Physics*, vol. 21, no. 12, pp. 905–940, 1982.

[32] K. E. Drexler, *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. Wiley, January 1992.

[33] J. M. Smith and E. Szathmary, *The Major Transitions in Evolution*. Oxford University Press, 1998.

[34] M. P. Fewell, "The atomic nuclide with the highest mean binding energy," *American Journal of Physics*, vol. 63, no. S2, 1995.

[35] F. J. Tipler, *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead*. Anchor, September 1997.

[36] G. Miller, *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. Anchor, 2001.

[37] A. Zahavi and A. Zahavi, *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press, 1999.

[38] K. E. Stanovich, *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. University of Chicago Press, new ed ed., October 2005.

[39] M. J. Baldwin, "A new factor in evolution," *The American Naturalist*, vol. 30, pp. 441–451, June 1896.

[40] W. Poundstone, *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. Contemporary Books, September 1985.

[41] E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning Ways for Your Mathematical Plays*, vol. 4. AK Peters, second revised edition ed., 2004.

[42] B. Franklin, *The Autobiography and Other Writings*. Penguin Classics, April 2003.

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

[44] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[45] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[46] R. Bellman, *Adaptive Control Processes*. Princeton University Press, 1961.

[47] J. Green, "Making book against oneself, the independence axiom, and non-linear utility theory," *Quarterly Journal of Economics*, vol. 98, pp. 785–796, 1987.

[48] D. Gillies, *Philosophical Theories of Probability*. Philosophical Issues in Science, Routledge, October 2000.