# One Decade of Universal Artificial Intelligence

**Marcus Hutter**

RSCS @ ANU and SML @ NICTA
Canberra, ACT, 0200, Australia

&

Department of Computer Science
ETH Zürich, Switzerland

February 2012

**Abstract**

The first decade of this century has seen the nascency of the first mathematical theory of general artificial intelligence. This theory of Universal Artificial Intelligence (UAI) has made significant contributions to many theoretical, philosophical, and practical AI questions. In a series of papers culminating in book (Hutter, 2005), an exciting sound and complete mathematical model for a super intelligent agent (AIXI) has been developed and rigorously analyzed. While nowadays most AI researchers avoid discussing intelligence, the award-winning PhD thesis (Legg, 2008) provided the philosophical embedding and investigated the UAI-based universal measure of rational intelligence, which is formal, objective and non-anthropocentric. Recently, effective approximations of AIXI have been derived and experimentally investigated in JAIR paper (Veness et al. 2011). This practical breakthrough has resulted in some impressive applications, finally muting earlier critique that UAI is only a theory. For the first time, without providing any domain knowledge, the same agent is able to self-adapt to a diverse range of interactive environments. For instance, AIXI is able to *learn* from scratch to play TicTacToe, Pacman, Kuhn Poker, and other games by trial and error, without even providing the rules of the games.

These achievements give new hope that the grand goal of Artificial General Intelligence is not elusive.

This article provides an informal overview of UAI in context. It attempts to gently introduce a very theoretical, formal, and mathematical subject, and discusses philosophical and technical ingredients, traits of intelligence, some social questions, and the past and future of UAI.

**Contents**




**Keywords**

artificial intelligence; reinforcement learning; algorithmic information theory; sequential decision theory; universal induction; rational agents; foundations.


*"The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science."*

*— Albert Einstein (1879–1955)*

# 1 Introduction

**The dream.** The *human mind* is one of the great mysteries in the Universe, and arguably the most interesting phenomenon to study. After all, it is connected to *consciousness* and *identity* which define who we are. Indeed, a healthy mind (and body) is our most precious possession. Intelligence is the most distinct characteristic of the human mind, and one we are particularly proud of. It enables us to understand, explore, and considerably shape our world, including ourselves. The field of *Artificial Intelligence* (AI) is concerned with the study and construction of artifacts that exhibit intelligent behavior, commonly by means of computer algorithms. The *grand goal* of AI is to develop systems that exhibit *general intelligence* on a *human-level* or beyond. If achieved, this would have a far greater impact on human society than all previous inventions together, likely resulting in a post-human civilization that only faintly resembles current humanity [Kur05, Hut12].

The dream of creating such artificial devices that reach or outperform our own intelligence is an old one with a persistent great divide between "optimists" and "pessimists". Apart from the overpowering technical challenges, research on machine intelligence also involves many fundamental philosophical questions with possibly inconvenient answers: What is intelligence? Can a machine be intelligent? Can a machine have free will? Does a human have free will? Is intelligence just an emergent phenomenon of a simple dynamical system or is it something intrinsically complex? What will our "Mind Children" be like? How does mortality affect decisions and actions? to name just a few.

**What was wrong with last century's AI.** Some claim that AI has not progressed much in the last 50 years. It definitely has progressed much slower than the fathers of AI expected and/or promised. There are also some philosophical arguments that the grand goal of creating super-human AI may even be elusive in principle. Both reasons have lead to a decreased interest in funding and research on the foundations of Artificial General Intelligence (AGI).

The real problem in my opinion is that early on, AI has focussed on the wrong paradigm, namely deductive logical; and being unable to get the foundations right in this framework, AI soon concentrated on practical but limited algorithms. Some prominent early researchers such as Ray Solomonoff, who actually participated in the 1956 Dartmouth workshop, generally regarded as the birth of AI, and later Peter Cheeseman and others, advocated a probabilistic inductive approach but couldn't compete with the soon dominating figures such as Marvin Minsky, Nils Nilsson, and others who advocated a symbolic/logic approach as the foundations of AI. (of course this paragraph is only a caricature of AI history).

Indeed it has even become an acceptable attitude that general intelligence is in principle unamenable to a formal definition. In my opinion, claiming something to be impossible without strong evidence sounds close to an unscientific position; and there *are no* convincing arguments against the feasibility of AGI [Cha96, Leg08].

Also, the failure of once-thought-promising AI-paradigms at best shows that they were not the right approach or maybe they only lacked sufficient computing power at the time. Indeed, after early optimism mid-last century followed by an AI depression, there is renewed, justified, optimism [RN10, Sec.1.3.10], as is evident by the new conference series on Artificial General Intelligence, the Blue Brain project, the Singularity movement, and the anthologies [GP07, WG12] prove. AI research has come in waves and paradigms (computation, logic, expert systems, neural nets, soft approaches, learning, probability). Finally, with the free access to unlimited amounts of data on the internet, *information*-centered AI research has blossomed.

**New foundations of A(G)I.** Universal Artificial Intelligence (UAI) is such a modern information-theoretic inductive approach to AGI, in which logical reasoning plays no direct role. UAI is a new paradigm to AGI via a path from universal induction to prediction to decision to action. It has been investigated in great technical depth [Hut05] and has already spawned promising formal definitions of rational intelligence, the optimal rational agent AIXI and practical approximations thereof, and put AI on solid mathematical foundations. It seems that we could, for the first time, have a general mathematical theory of (rational) intelligence that is sound and complete in the sense of well-defining the general AI problem as detailed below. The theory allows a rigorous mathematical investigation of many interesting philosophical questions surrounding (artificial) intelligence. Since the theory is complete, definite answers can be obtained for a large variety of intelligence-related questions, as foreshadowed by the award winning PhD thesis of [Leg08].

**Contents.** Section 2 provides the context and background for UAI. It will summarize various last century's paradigms for and approaches to understanding and

building artificial intelligences, highlighting their problems and how UAI is similar or different to them. Section 3 then informally describes the ingredients of UAI. It mentions the UAI-based intelligence measure only in passing to go directly to the core AIXI definition. In which sense AIXI is the most intelligent agent and a theoretical solution of the AI problem is explained. Section 4 explains how the complex phenomenon of intelligence with all its facets can emerge from the simple AIXI equation. Section 5 considers an embodied version of AIXI embedded into our society. I go through some important social questions and hint at how AIXI might behave, but this is essentially unexplored terrain. The technical state-of-the-art/development of UAI is summarized in Section 6: theoretical results for AIXI and universal Solomonoff induction; practical approximations, implementations, and applications of AIXI; UAI-based intelligence measures, tests, and definitions; and the human knowledge compression contest. Section 7 concludes with a summary and outlook how UAI helps in formalizing and answering deep philosophical questions around AGI and last but not least how to build super intelligent agents.

# 2 The AGI Problem

The term AI means different things to different people. I will first discuss why this is so, and will argue that this due to a lack of solid and generally agreed-upon foundations of AI. The field of AI soon abandoned its efforts of rectifying this state of affairs, and pessimists even created a defense mechanism denying the possibility or usefulness of a (simple) formal theory of general intelligence. While human intelligence might indeed be messy and unintelligible, I will argue that a simple formal definition of machine intelligence *is* possible and useful. I will discuss how this definition fits into the various important dimensions of research on (artificial) intelligence including human↔rational, thinking↔acting, top-down↔bottom-up, the agent framework, traits of intelligence, deduction↔induction, and learning↔planning.

**The problem.** I define *the AI problem* to mean the problem of building systems that possess general, rather than specific, intelligence in the sense of being able to solve a wide range of problems generally regarded to require human-level intelligence.

Optimists believe that the AI problem can be solved within a couple of decades [Kur05]. Pessimists deny its principle feasibility on religious, philosophical, mathematical, or technical grounds (see [RN10, Chp.26] for a list of arguments). Optimists have refuted/rebutted all those arguments (see [Cha96, Chp.9] and [Leg08]), but haven't produced super-human AI either, so the issue remains unsettled.

One problem in AI, and I will argue key problem, is that there is no general agreement on what intelligence is. This has lead to endless circular and often fruitless arguments, and has held up progress. Generally, the lack of a generally-accepted solid foundation makes high card houses fold easily. Compare this with Russell's paradox which shattered the foundations of mathematics, and which was finally resolved by the completely formal and generally agreed-upon ZF(C) theory of sets.

On the other hand, it is an anomaly that nowadays most AI researchers avoid

discussing or formalizing intelligence, which is caused by several factors: It is a difficult old subject, it is politically charged, it is not necessary for narrow AI which focusses on specific applications, AI research is done primarily by computer scientists who mainly care about algorithms rather than philosophical foundations, and the popular belief that general intelligence is principally unamenable to a mathematical definition. These reasons explain but only partially justify the limited effort in trying to formalize general intelligence. There is no convincing argument that this is impossible.

Assume we had a formal, objective, non-anthropocentric, and direct definition, measure, and/or test of intelligence, or at least a very general intelligence-resembling formalism that could serve as an adequate substitute. This would bring the higher goals of the field into tight focus and allow us to objectively and rigorously compare different approaches and judge the overall progress. Formalizing and rigorously defining a previously vague concept usually constitutes a quantum leap forward in the field: Cf. the history of sets, numbers, logic, fluxions/infinitesimals, energy, infinity, temperature, space, time, observer, etc.

Is a simple *formal definition of intelligence* possible? Isn't intelligence a too complex and anthropocentric phenomenon to allow formalization? Likely not: There are very simple models of chaotic phenomena such as turbulence. Think about the simple iterative map $z \to z^2 + c$ that produces the amazingly rich, fractal landscape, sophisticated versions of it used to produce images of virtual ecosystems as in the movie Avatar. Or the complexity of (bio)chemistry emerges out of the elegant mathematical theory Quantum Electro Dynamics.

Modeling human intelligence is probably going to be messy, but ideal rational behavior seems to capture the essence of intelligence, and, as I claim, can indeed be completely formalized. Even if there is no unique definition capturing all aspects we want to include in a definition of intelligence, or if some aspects are forever beyond formalization (maybe consciousness and qualia), pushing the frontier and studying the best available formal proxy is of utmost importance for understanding artificial and natural minds.

**Context.** There are many fields that try to understand the phenomenon of intelligence and whose insights help in creating intelligent systems: *cognitive psychology* [SMM07] and *behaviorism* [Ski74], *philosophy of mind* [Cha02, Sea05], *neuroscience* [HB04], *linguistics* [Hau01, Cho06], *anthropology* [Par07], *machine learning* [SB98, Bis06], *logic* [Tur84, Llo87], *computer science* [RN10], *biological evolution* [TTJ01, Kar07], *economics* [McK09], and *others*.

| **What is AI?** | **Thinking** | **Acting** |
|---|---|---|
| **humanly** | Cognitive Science | Turing test, Behaviorism |
| **rationally** | Laws of Thought | **Doing the Right Thing** |

Cognitive science studies how humans think, Behaviorism and the Turing test how humans act, the laws of thought define rational thinking, while AI research increasingly focusses on systems that act rationally.

In computer science, most AI research is *bottom-up*; extending and improving existing or developing new *algorithms* and increasing their range of applicability;

an interplay between experimentation on toy problems and theory, with occasional real-world applications. A *top-down* approach would start from a general principle and derive effective approximations (like heuristic approximations to minimax tree search). Maybe when the top-down and bottom-up approaches meet in the middle, we will have arrived at practical truly intelligent machines.

The science of artificial intelligence may be defined as the construction of intelligent systems (*artificial agents*) and their analysis. A natural definition of a *system* is anything that has an input and an output stream, or equivalently an agent that acts and observes. This agent perspective of AI [RN10] brings some order and unification into the large variety of problems the fields wants to address, but it is only a framework rather than providing a complete theory of intelligence. In the absence of a perfect (stochastic) model of the environment the agent interacts with, *machine learning* techniques are needed and employed to learn from experience. There is no general theory for learning agents (apart from UAI). This has resulted in an ever increasing number of *limited models and algorithms* in the past.

What distinguishes an *intelligent* system from a non-intelligent one? *Intelligence* can have many faces like *reasoning*, *creativity*, *association*, *generalization*, *pattern recognition*, *problem solving*, *memorization*, *planning*, *achieving goals*, *learning*, *optimization*, *self-preservation*, *vision*, *language processing*, *classification*, *induction*, *deduction*, and *knowledge acquisition and processing*. A formal definition incorporating every aspect of intelligence, however, *seems* difficult.

There is no lack of attempts to characterize or define intelligence trying to capture all traits *informally* [LH07a]. One of the more successful characterizations is: *Intelligence measures an agents ability to perform well in a large range of environments* [LH07b]. Most traits of intelligence are implicit in and emergent from this definition as these capacities enable an agent to succeed [Leg08]. Convincing formal definitions other than the ones spawned by UAI are essentially lacking.

Another important dichotomy is whether an approach focusses (more) on deduction or induction. Traditional AI concentrates mostly on the logical deductive reasoning aspect, while machine learning focusses on the inductive inference aspect. Learning and hence induction are indispensable traits of any AGI. Regrettably, induction is peripheral to traditional AI, and the machine learning community in large is not interested in A(G)I. It is the field of reinforcement learning at the intersection of AI and machine learning that has AGI ambitions *and* takes learning seriously.

**UAI in perspective.** The theory of Universal Artificial Intelligence developed in the last decade is a modern information-theoretic, inductive, reinforcement learning approach to AGI that has been investigated in great technical depth [Hut05].

Like traditional AI, UAI is concerned with agents *doing the right thing*, but is otherwise quite different: It is a *top-down* approach in the sense that it starts with a single completely *formal general* definition of intelligence from which an essentially *unique agent* that seems to possess all *traits* of rational intelligence is derived. It is not just another framework with some gaps to be filled in later, since the agent is *completely* defined.

It also takes induction very seriously: Universal learning is one of the agent's

two key elements (the other is stochastic planning). Indeed, logic and deduction play no fundamental role in UAI (but are emergent). This also naturally dissolves Lucas' and Penrose' [Pen94] argument against AGI that Goedel's incompleteness result shows that the human mind is not a computer. The fallacy is to assume that the mind (human and machine alike) are infallible deductive machines.

The status of UAI might be compared to Super String theory in physics. Both are currently the most promising candidates for a grand unification (of AI and physics, respectively), although there are also marked differences. Like the unification hierarchy of physical theories allows relating and regarding the myriad of limited models as effective approximations, UAI allows us to regard existing approaches to AI as effective approximations. Understanding AI in this way gives researchers a much more coherent view of the field.

Indeed, UAI seems to be the first sound and complete mathematical theory of (rational) intelligence. The next section presents a very brief introduction to UAI from [Hut09c], together with an informal explanation of what the previous sentence actually means. See [Hut05] for formal definitions and results.

# 3 Universal Artificial Intelligence

This section describes the theory of Universal Artificial Intelligence (UAI), a modern information-theoretic approach to AI, which differs essentially from mainstream A(G)I research described in the previous sections. The connection of UAI to other research fields and the philosophical and technical ingredients of UAI (Ockham, Epicurus, Turing, Bayes, Solomonoff, Kolmogorov, Bellman) are briefly discussed. The UAI-based universal intelligence measure and order relation in turn define the (w.r.t. this measure) most intelligent agent AIXI, which seems to be the first sound and complete theory of a universal optimal rational agent embedded in an arbitrary computable but unknown environment with reinforcement feedback. The final paragraph clarifies what this actually means.

**Defining Intelligence.** Philosophers, AI researchers, psychologists, and others have suggested many informal=verbal definitions of intelligence [LH07a], but there is not too much work on formal definitions that are broad, objective, and non-anthropocentric. See [LH07b] for a comprehensive collection, discussion and comparison of intelligence definitions, tests, and measures with all relevant references. It is beyond the scope of this article to discuss them.

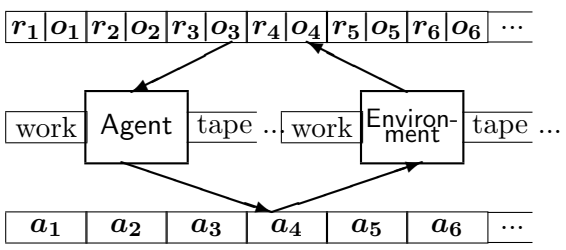
Intelligence is graded, since agents can be more or less intelligent. Therefore it is more natural to consider measures of intelligence, rather than binary definitions which would classify agents as intelligent or not based on an (arbitrary) threshold. This is exactly what UAI provides: A formal, broad, objective, universal measure of intelligence [LH07b], which formalizes the verbal characterization stated in the previous section. Agents can be more or less intelligent w.r.t. this measure and hence can be sorted w.r.t. their intelligence [Hut05, Sec.5.1.4]. One can show that there is an agent, coined AIXI, that maximizes this measure, which could therefore

be called the most intelligent agent.

I will not present the UAI-based intelligence measure [LH07b] and order relation [Hut05] here, but, after listing the conceptual ingredients to UAI and AIXI, directly proceed to defining and discussing AIXI.

**UAI and AIXI ingredients [Hut09c].** The theory of UAI has interconnections with (draws from and contributes to) many research fields, encompassing computer science (artificial intelligence, machine learning, computation), engineering (information theory, adaptive control), economics (rational agents, game theory), mathematics (statistics, probability), psychology (behaviorism, motivation, incentives), and philosophy (inductive inference, theory of knowledge). The concrete ingredients in AIXI are as follows: Intelligent *actions* are based on informed *decisions*. Attaining good decisions requires *predictions* which are typically based on models of the environments. Models are constructed or learned from past observations via *induction*. Fortunately, based on the *deep philosophical insights* and *powerful mathematical developments*, all these problems have been overcome, at least in theory: So what do we need (from a mathematical point of view) to construct a universal optimal learning agent interacting with an arbitrary unknown environment? The theory, coined *UAI*, developed in the last decade and explained in [Hut05] says: *All you need is Ockham*, *Epicurus*, *Turing*, *Bayes*, *Solomonoff* [Sol64], *Kolmogorov* [Kol65], and *Bellman* [Bel57]: Sequential decision theory [Ber06b] (*Bellman*'s equation) formally solves the problem of rational agents in uncertain worlds if the true environmental probability distribution is known. If the environment is unknown, *Bayes*ians [Ber93] replace the true distribution by a weighted mixture of distributions from some (hypothesis) class. Using the large class of all (semi)measures that are (semi)computable on a *Turing* machine bears in mind *Epicurus*, who teaches not to discard any (consistent) hypothesis. In order not to ignore *Ockham*, who would select the simplest hypothesis, *Solomonoff* defined a universal prior that assigns high/low prior weight to simple/complex environments [RH11], where *Kolmogorov* quantifies complexity [LV08]. Their unification constitutes the theory of UAI and resulted in the universal intelligence measure and order relation and the following model/agent AIXI.

**The AIXI Model in one line [Hut09c].** It is possible to write down the AIXI model explicitly in one line, although *one should not expect to be able to grasp the full meaning and power from this compact and somewhat simplified representation.*




AIXI is an agent that interacts with an environment in cycles $k = 1, 2, ..., m$. In cycle $k$, AIXI takes action $a_k$ (e.g. a limb movement) based on past perceptions $o_1 r_1 .. o_{k-1} r_{k-1}$ as defined below. Thereafter, the environment provides a (regular) observation $o_k$ (e.g. a camera image) to AIXI and a real-valued reward $r_k$. The reward can be very scarce, e.g. just +1 (-1) for winning (losing) a chess game, and 0 at all other times. Then the

next cycle $k+1$ starts. This agent-environment interaction protocol can be depicted as on the right. Given the interaction protocol above, the simplest version of AIXI is defined by

$$\text{AIXI} \qquad a_k \;:=\; \arg\max_{a_k} \sum_{o_k r_k} ... \max_{a_m} \sum_{o_m r_m} [r_k + ... + r_m] \sum_{q\,:\,U(q,a_1..a_m)=o_1 r_1..o_m r_m} 2^{-\ell(q)}$$

The expression shows that AIXI tries to maximize its total future reward $r_k + ... + r_m$. If the environment is modeled by a deterministic program $q$, then the future perceptions $...o_k r_k..o_m r_m = U(q, a_1..a_m)$ can be computed, where $U$ is a universal (monotone Turing) machine executing $q$ given $a_1..a_m$. Since $q$ is unknown, AIXI has to maximize its expected reward, i.e. average $r_k + ... + r_m$ over all possible future perceptions created by all possible environments $q$ that are consistent with past perceptions. The simpler an environment, the higher is its a-priori contribution $2^{-\ell(q)}$, where simplicity is measured by the length $\ell$ of program $q$. AIXI effectively learns by eliminating Turing machines $q$ once they become inconsistent with the progressing history. Since noisy environments are just mixtures of deterministic environments, they are automatically included [RH11, Sec.7.2],[WSH11]. The sums in the formula constitute the averaging process. Averaging and maximization have to be performed in chronological order, hence the interleaving of max and $\Sigma$ (similarly to minimax for games).

One can fix any finite action and perception space, any reasonable $U$, and any large finite lifetime $m$. This completely and uniquely defines AIXI's actions $a_k$, which are limit-computable via the expression above (all quantities are known).

**Discussion.** The AIXI model seems to be the first sound and complete *theory* of a universal optimal rational agent embedded in an arbitrary computable but unknown environment with reinforcement feedback. AIXI is *universal* in the sense that it is designed to be able to interact with any (deterministic or stochastic) computable environment; the universal Turing machines on which it is based is crucially responsible for this. AIXI is *complete* in the sense that it is not an incomplete framework or partial specification (like Bayesian statistics which leaves open the choice of the prior or the rational agent framework or the subjective expected utility principle) but is completely and essentially uniquely defined. AIXI is *sound* in the sense of being (by construction) free of any internal contradictions (unlike e.g. in knowledge-based deductive reasoning systems where avoiding inconsistencies can be very challenging). AIXI is *optimal* in the senses that: no other agent can perform uniformly better or equal in all environments, it is a unification of two optimal theories themselves, a variant is self-optimizing; and it is likely also optimal in other/stronger senses. AIXI is *rational* in the sense of trying to maximize its future long-term reward. For the reasons above I have argued that AIXI is a mathematical "solution" of the AI problem: AIXI would be able to learn any learnable task and likely better so than any other unbiased agent, but AIXI is more a *theory* or formal definition rather than an algorithm, since it is only limit-computable. How can an equation that fits into a single line capture the diversity, complexity, and essence of (rational) intelligence? We know that complex appearing phenomena such as chaos and fractals

can have simple descriptions such as iterative maps and the complexity of chemistry emerges from simple physical laws. There is no a-priori reason why ideal rational intelligent behavior should not also have a simple description, with most traits of intelligence being emergent. Indeed, even an axiomatic characterization seems possible [SH11a, SH11b].

# 4 Facets of Intelligence

Intelligence can have many faces. I will argue in this section that the AIXI model possesses all or at least most properties an intelligent rational agent should possess. Some facets have already been formalized, some are essentially built-in, but the majority have to be emergent. Some of the claims have been proven in [Hut05] but the majority has yet to be addressed.

**Generalization** is essentially inductive inference [RH11]. ***Induction*** is the process of inferring general laws or models from observations or data by finding regularities in past/other data. This trait is a fundamental cornerstone of intelligence.

**Prediction** is concerned with forecasting future observations (often based on models of the world learned) from past observations. Solomonoff's theory of prediction [Sol64, Sol78] is a universally optimal solution of the prediction problem [Hut07, RH11]. Since it is a key ingredient in the AIXI model, it is natural to expect that AIXI is an optimal predictor if rewarded for correct predictions. Curiously only weak and limited rigorous results could be proven so far [Hut05, Sec.6.2].

**Pattern recognition**, abstractly speaking, is concerned with classifying data (patterns). This requires a similarity measure between patterns. Supervised ***classification*** can essentially be reduced to a sequence prediction problem, hence formally pattern recognition reduces to the previous item, although interesting questions specific to classification emerge [Hut05, Chp.3].

**Association.** Two stimuli or observations are associated if there exists some (cor)relation between them. A set of observations can often be ***clustered*** into different categories of similar=associated items. For AGI, a *universal* similarity measure is required. Kolmogorov complexity via the universal similarity metric [CV05] can provide such a measure, but many fundamental questions have yet to be explored: How does association function in AIXI? How can Kolmogorov complexity well-define the (inherently? so far?) ill-defined clustering problem?

**Reasoning** is arguably the most prominent trait of human intelligence. Interestingly deductive reasoning and logic are ***not*** part of the AIXI architecture. The fundamental assumption is that there is no sure knowledge of the world, all inference is tentative and inductive, and that logic and ***deduction*** constitute an idealized limit applicable in situations where uncertainties are extremely small, i.e. probabilities are extremely close to 1 or 0. What would be very interesting to show is that ***logic*** is an emergent phenomenon, i.e. that AIXI learns to reason logically if/since this helps collect reward.

**Problem solving** might be defined as goal-oriented reasoning, and hence reduces to the previous item, since AIXI is designed to ***achieve goals*** (which is reward maximization in the special case of a terminal reward when the goal is achieved). Problems can be of very different nature, and some of the other traits of intelligence can be regarded as instances of problem solving, e.g. planning.

**Planning** ability is directly incorporated in AIXI via the alternating maximization and summation in the definition. Algorithmically AIXI plans through its entire life via a deep expectimax tree search up to its death, based on its belief about the world. In known constrained domains this search corresponds to classical exact planning strategies as e.g. exemplified in [Hut05, Chp.6].

**Creativity** is the ability to generate innovative ideas and to manifest these into reality. Creative people are often more successful than unimaginative ones. Since AIXI is the ultimate success-driven agent, AIXI should be highly creative, but this has yet to be formalized and proven, or at least exemplified.

**Knowledge.** AIXI stores the entire interaction history and has perfect ***memory***. Additionally, models of the experienced world are constructed (learned) from this ***information*** in form of short(est) programs. These models guide AIXI's behavior, so constitute knowledge for AIXI. Any ***ontology*** is implicit in these programs. How short-term, long-term, relational, hierarchical, etc. memory emerges out of this compression-based approach has not yet been explored.

**Actions** influence the environment which reacts back to the agent. ***Decisions*** can have long-term consequences, which the expectimax planner of AIXI should properly take into account. Particular issues of concern are the interplay of learning and planning (the infamous exploration↔exploitation tradeoff [LH11]). Additional complications that arise from embodied agents will be considered in the next section.

**Learning.** There are many different forms of learning: supervised, unsupervised, semi-supervised, reinforcement, transfer, associative, transductive, prequential, and many others. By design, AIXI is a reinforcement learner, but one can show that it will also "listen" to an informative teacher, i.e. it *learns* to learn supervised [Hut05, Sec.6.5]. It is plausible that AIXI can also acquire the other learning techniques.

**Self-awareness** allows one to (meta)reason about one's own thoughts, which is an important trait of higher intelligence, in particularly when interacting with other forms of intelligence. Technically all what might be needed is that an agent has and exploits not only a model of the world but also a model of itself including aspects of its own algorithm, and this recursively. Is AIXI self-aware in this technical sense?

**Consciousness** is possibly the most mysterious trait of the human mind. Whether anything rigorous can ever be said about the consciousness of AIXI or AIs in general is not clear and in any case beyond my expertise. I leave this to philosophers of the mind [Cha02] like the world-renowned expert on (the hard problem of) consciousness, David Chalmers [Cha96].

# 5 Social Questions

Consider now a sophisticated physical humanoid robot like Honda's ASIMO but equipped with an AIXI brain. The observations $o_k$ consist of camera image, microphone signal, and other sensory input. The actions $a_k$ consist of controlling mainly a loud speaker and motors for limbs, but possibly other internal functions it has direct control over. The reward $r_k$ should be some combination of its own "well-being" (e.g. proportional to its battery level and condition of its body parts) and external reward/punishment from some "teacher(s)".

Imagine now what happens if this AIXI-robot is let loose in our society. Many questions deserving attention arise, and some are imperative to be rigorously investigated before risking this experiment.

Children of higher animals require extensive nurturing in a safe environment because they lack sufficient innate skills for survival in the real world, but are compensated for their ability to learn to perform well in a large range of environments. AIXI is at the extreme of being "born" with essentially no knowledge about our world, but a universal "brain" for learning and planning in any environment where this is possible. As such, it also requires a guiding teacher initially. Otherwise it would simply run out of battery.

AIXI has to learn *vision*, *language*, and *motor skills* from scratch, similarly to higher animals and machine learning algorithms, but more extreme/general. Indeed, Solomonoff [Sol64] already showed how his system can learn grammar from positive instances only, but much remains to be done. Appropriate *training sequences* and *reward shaping* in this early "childhood" phase of AIXI are important. AIXI can learn from rather crude teachers as long as the reward is biased in the 'right' direction. The answers to many of the following questions likely depend on the upbringing of AIXI:

- ***Schooling:*** Will a pure reward maximizer such as AIXI listen to and trust a teacher and learn to learn supervised (=faster)? Yes [Hut05, Sec.6.5].
- Take ***Drugs*** (hacking the reward system): Likely no, since long-term reward would be small (death), but see [RO11].
- ***Replication or procreation:*** Likely yes, if AIXI believes that clones or descendants are useful for its own goals.
- ***Suicide:*** Likely yes (no), if AIXI is raised to believe to go to heaven (hell) i.e. maximal (minimal) reward forever.
- ***Self-Improvement:*** Likely yes, since this helps to increase reward.
- ***Manipulation:*** Manipulate or threaten teacher to give more reward.
- ***Attitude:*** Are pure reward maximizers egoists, *psychopaths*, and/or killers or will they be *friendly* (*altruism* as extended *ego(t)ism*)?
- ***Curiosity*** killed the cat and maybe AIXI, or is extra reward for curiosity necessary [Sch07, Ors10]?
- ***Immortality*** can cause laziness [Hut05, Sec.5.7]!
- Can ***self-preservation*** be learned or need (parts of) it be innate.

- ***Socializing:*** How will AIXI interact with another AIXI [Hut09c, Sec.5j],[PH06]?

A partial discussion of some of these questions can be found in [Hut05] but many are essentially unexplored. Point is that since AIXI is completely formal, it permits to formalize these questions and to mathematically analyze them. That is, UAI has the potential to arrive at definite answers to various questions regarding the social behavior of super-intelligences. Some formalizations and semi-formal answers have recently appeared in the award-winning papers [OR11, RO11].

# 6 State of the Art

This section describes the technical achievements of UAI to date. Some remarkable and surprising results have already been obtained. Various theoretical consistency and optimality results for AIXI have been proven, although stronger results would be desirable. On the other hand, the special case of universal induction and prediction in non-reactive environments is essentially closed. From the practical side, various computable approximations of AIXI have been developed, with the latest MC-AIXI-CTW incarnation exhibiting impressive performance. Practical approximations of the universal intelligence measure have also been used to test and consistently order systems of limited intelligence. Some other related work such as the compression contest is also briefly mentioned, and references to some more practical but less general work such as feature reinforcement learning are given.

**Theory of UAI.** Forceful theoretical arguments that AIXI is the most intelligent general-purpose agent incorporating all aspects of rational intelligence have been put forward, supported by partial proofs. For this, results of many fields had to be pulled together or developed in the first place: *Kolmogorov complexity* [LV08], *information theory* [CT06], *sequential decision theory* [Ber06b], *reinforcement learning* [SB98], *artificial intelligence* [RN10], *Bayesian statistics* [Ber06a], *universal induction* [RH11], and *rational agents* [SLB09]. Various notions of optimality have been considered. The difficulty is coming up with sufficiently strong but still satisfiable notions. Some are weaker than desirable, others are too strong for any agent to achieve. What has been shown thus far is that AIXI learns the correct predictive model [Hut05], is Pareto optimal in the sense that no other agent can perform uniformly better or equal in all environments, and a variant is self-optimizing in the sense that asymptotically the accumulated reward is as high as possible, i.e. the same as the maximal reward achievable by a completely informed agent [Hut02b]. AIXI is likely also optimal in other/stronger senses. An axiomatic characterization has also been developed [SH11a, SH11b].

**The induction problem.** The induction problem is a fundamental problem in philosophy [Ear93, RH11] and science [Jay03, Wal05, GHW11], and a key subcomponent of UAI. Classical open problems around induction are the zero prior problem and the confirmation of (universal) hypotheses in general and the Black

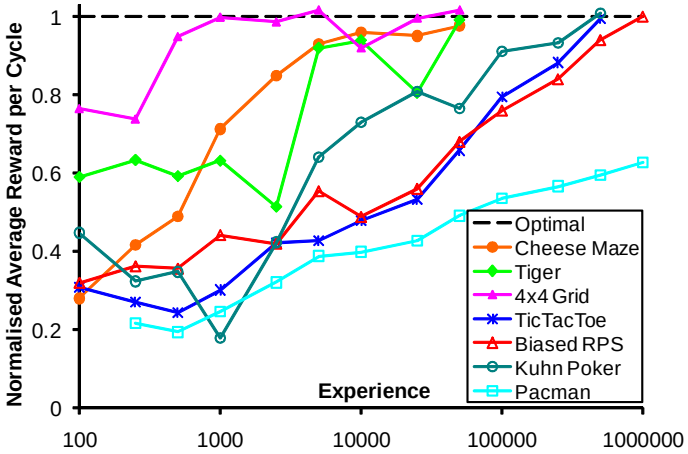
ravens paradox in particular, reparametrization invariance, the old-evidence problem and ad-hoc hypotheses, and the updating problem [Ear93]. In a series of papers (see [Hut07] for references) it has been shown that Solomonoff's theory of universal induction essentially solves or circumvents all these problems [RH11]. It is also predictively optimal and has minimal regret for arbitrary loss functions.

It is fair to say that Solomonoff's theory serves as an adequate mathematical/theoretical foundation of induction [RH11], machine learning [Hut11], and component of UAI [Hut05].

**Computable approximations of AIXI.** An early critique of UAI was that AIXI is incomputable. The down-scaled still provably optimal AIXI*tl* model [Hut05, Chp.7] based on universal search algorithms [Lev73, Hut02a, Gag07] was still computationally intractable. The Optimal Ordered Problem Solver [Sch04] was the first practical implementation of universal search and has been able to solve open learning tasks such as Towers-of-Hanoi for arbitrary number of disks, robotic behavior, and others.

For repeated $2 \times 2$ matrix games such as the Prisoner's dilemma, a direct brute-force approximation of AIXI is computationally tractable. Despite these domains being tiny, they raise notoriously difficult questions [SLB09]. The experimental results confirmed the theoretical optimality claims of AIXI [PH06], as far as limited experiments are able to do so.

A Monte-Carlo approximation of AIXI has been proposed in [Pan08] that samples programs according to their algorithmic probability as a way of approximating Solomonoff's universal a-priori probability, similar to sampling from the speed prior [Sch02].




The most powerful systematic approximation, implementation, and application of AIXI so far is the MC-AIXI-CTW algorithm [VNHS10]. It combines award-winning ideas from universal Bayesian data compression [WST95] and the recent highly successful (in computer Go) upper confidence bound algorithm for expectimax tree search [KS06]. For the first time, without any domain knowledge, the same agent is able to self-adapt to a diverse range of environments. For instance, AIXI, is able to *learn* from scratch how to play TicTacToe, Pacman, Kuhn Poker, and other games by trial and error without even providing the rules of the games [VNH+11].

**Measures/tests/definitions of intelligence.** The history of informal definitions and measures of intelligence [LH07a] and anthropocentric tests of intelligence [Tur50] is long and old. In the last decade various formal definitions, mea-

sures and tests have been suggested: Solomonoff induction and Kolmogorov complexity inspired the universal C-test [HO00, HOMC98], while AIXI inspired an extremely general, objective, fundamental, and formal intelligence order relation [Hut05] and a universal intelligence measure [LH07b, Leg08], which have already attracted the popular scientific press [Fié05] and received the SIAI award. Practical instantiations thereof [HOD10, LV11] also received quite some media attention (http://users.dsic.upv.es/proy/anynt/).

**Less related/general work.** There is of course other less related, less general work, similar in spirit to or with similar aims as UAI/AIXI, e.g. UTree [McC96], URL [FMRW10], PORL [SHL97, SH99], FOMDP [SB09], FacMDP [SDL07], PSR [SLJ+03], POMDP [DV09], and others. The feature reinforcement learning approach also belongs to this category [Hut09b, Hut09a, SH10, NSH11].

**Compression contest.** The ongoing Human Knowledge Compression Contest [Hut06] is another outgrowth of UAI. The contest is motivated by the fact that being able to compress well is closely related to being able to predict well and ultimately to act intelligently, thus reducing the slippery concept of intelligence to hard file size numbers. Technically it is a community project to approximate Kolmogorov complexity on real-world textual data. In order to compress data, one has to find regularities in them, which is intrinsically difficult (many researchers live from analyzing data and finding compact models). So compressors better than the current "dumb" compressors need to be smart(er). Since the prize wants to stimulate the development of "universally" smart compressors, a "universal" corpus of data has been chosen. Arguably the online encyclopedia Wikipedia is a good snapshot of the Human World Knowledge. So the ultimate compressor of it should "understand" all human knowledge, i.e. be really smart. The contest is meant to be a cost-effective way of motivating researchers to spend time towards achieving AGI via the promising and quantitative path of compression. The competition raised considerable attention when launched, but to retain attention the prize money should be increased (sponsors are welcome), and the setup needs some adaptation.

# 7 Discussion

**Formalizing and answering deep philosophical questions.** UAI deepens our understanding of artificial (and to a limited extent human) intelligence; in particular which and how facets of intelligence can be understood as emergent phenomena of goal- or reward-driven actions in unknown environments. UAI allows a more quantitative and rigorous discussion of various philosophical questions around intelligence, and ultimately settling these questions. This can and partly has been done by formalizing the philosophical concepts related to intelligence under consideration, and by studying them mathematically. Formal definitions may not perfectly or not one-to-one or not uniquely correspond to their intuitive counterparts, but in this case alternative formalizations allow comparison and selection. In this way it might

even be possible to rigorously answer various social and ethical questions: whether a super rational intelligence such as AIXI will be benign to humans and/or its ilk, or behave psychopathically and kill or enslave humans, or be insane and e.g. commit suicide.

**Building more intelligent agents.** From a practical point of building intelligent agents, since AIXI is incomputable or more precisely only limit-computable, it has to be approximated in practice. The results achieved with the MC-AIXI-CTW approximation are only the beginning. As outlined in [VNH+11], many variations and extensions are possible, in particular to incorporate long-term memory and smarter planning heuristics. The same single MC-AIXI-CTW agent is already able to learn to play TicTacToe, Kuhn Poker, and most impressively Pacman [VNH+11] from scratch. Besides Pacman, there are hundreds of other arcade games from the 1980s, and it would be sensational if a single algorithm could learn them all solely by trial and error, which seems feasible for (a variant of) MC-AIXI-CTW. While these are "just" recreational games, they *do* contain many prototypical elements of the real world, such as food, enemies, friends, space, obstacles, objects, and weapons. Next could be a test in modern virtual worlds (e.g. bots for VR/role games or intelligent software agents for the internet) that require intelligent agents, and finally some selected real-world problems.

**Epilogue.** It is virtually impossible to predict the future rate of progress but past progress on UAI makes me confident that UAI as a whole will continually progress. By providing rigorous foundations to AI, I believe that UAI will also speed up progress in the field of A(G)I in general. In any case, UAI is a very useful educational tool with AIXI being a gold standard for intelligent agents which other practical general purpose AI programs should aim for.

# References


[Bel57] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[Ber93] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, Berlin, 3rd edition, 1993.

[Ber06a] J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[Ber06b] D. P. Bertsekas. *Dynamic Programming and Optimal Control, volume I and II*. Athena Scientific, Belmont, MA, 3rd edition, 2006. Volumes 1 and 2.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Cha96] D. J. Chalmers. *The Conscious Mind*. Oxford University Press, USA, 1996.

[Cha02] D. J. Chalmers, editor. *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press, USA, 2002.

[Cho06] N. Chomsky. *Language and Mind*. Cambridge University Press, 3rd edition, 2006.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Intersience, 2nd edition, 2006.

[CV05] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Trans. Information Theory*, 51(4):1523–1545, 2005.

[DV09] Finale Doshi-Velez. The infinite partially observable markov decision process. In *Proc. 22nd Conf. on Neural Information Processing Systems 22 (NIPS'09)*, 2009.

[Ear93] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, MA, 1993.

[Fié05] C. Fiévet. Mesurer l'intelligence d'une machine. In *Le Monde de l'intelligence*, volume 1, pages 42–45, Paris, November 2005. Mondeo publishing.

[FMRW10] V.F. Farias, C. C. Moallemi, B. Van Roy, and T. Weissman. Universal reinforcement learning. *IEEE Transactions on Information Theory*, 56(5):2441–2454, 2010.

[Gag07] M. Gaglio. Universal search. *Scholarpedia*, 2(11):2575, 2007.

[GHW11] D. M. Gabbay, S. Hartmann, and J. Woods, editors. *Handbook of Inductive Logic*. North Holland, 2011.

[GP07] B. Goertzel and C. Pennachin, editors. *Artificial General Intelligence*. Springer, 2007.

[Hau01] R. Hausser. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Springer, 2nd edition, 2001.

[HB04] J. Hawkins and S. Blakeslee. *On Intelligence*. Times Books, 2004.

[HO00] J. Hernández-Orallo. On the computational measurement of intelligence factors. In *Performance Metrics for Intelligent Systems Workshop*, pages 1–8, Gaithersburg, MD, USA, 2000.

[HOD10] J. Hernandez-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539, 2010.

[HOMC98] J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of kolmogorov complexity. In *International Symposium of Engineering of Intelligent Systems*, pages 146–163, 1998.

[Hut02a] M. Hutter. The fastest and shortest algorithm for all well-defined problems. *International Journal of Foundations of Computer Science*, 13(3):431–443, 2002.

[Hut02b] M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 364–379, Sydney, 2002. Springer, Berlin.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hut06] M. Hutter. Human knowledge compression prize, 2006. open ended, http://prize.hutter1.net/.

[Hut07] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

[Hut09a] M. Hutter. Feature dynamic Bayesian networks. In *Proc. 2nd Conf. on Artificial General Intelligence (AGI'09)*, volume 8, pages 67–73. Atlantis Press, 2009.

[Hut09b] M. Hutter. Feature reinforcement learning: Part I: Unstructured MDPs. *Journal of Artificial General Intelligence*, 1:3–24, 2009.

[Hut09c] M. Hutter. Open problems in universal induction & intelligence. *Algorithms*, 3(2):879–906, 2009.

[Hut11] M. Hutter. Universal learning theory. In C. Sammut and G. Webb, editors, *Encyclopedia of Machine Learning*, pages 1001–1008. Springer, 2011.

[Hut12] M. Hutter. Can intelligence explode? *Journal of Consciousness Studies*, 19(1-2):??–??, 2012.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, MA, 2003.

[Kar07] K. V. Kardong. *An Introduction to Biological Evolution*. McGraw-Hill Science/Engineering/Math, 2nd edition, 2007.

[Kol65] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):1–7, 1965.

[KS06] L. Kocsis and C. Szepesvári. Bandit based Monte-Carlo planning. In *Proc. 17th European Conf. on Machine Learning (ECML'06)*, pages 282–293, 2006.

[Kur05] R. Kurzweil. *The Singularity Is Near*. Viking, 2005.

[Leg08] S. Legg. *Machine Super Intelligence*. PhD thesis, IDSIA, Lugano, Switzerland, 2008. Recipient of the $10'000,- Singularity Prize/Award.

[Lev73] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.

[LH07a] S. Legg and M. Hutter. A collection of definitions of intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157 of *Frontiers in Artificial Intelligence and Applications*, pages 17–24, Amsterdam, NL, 2007. IOS Press.

[LH07b] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

[LH11] T. Lattimore and M. Hutter. Asymptotically optimal agents. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 368–382, Espoo, Finland, 2011. Springer, Berlin.

[Llo87] J. W. Lloyd. *Foundations of Logic Programming*. Springer, 2nd edition, 1987.

[LV08] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, Berlin, 3rd edition, 2008.

[LV11] S. Legg and J. Veness. An approximation of the universal intelligence measure. In *Proc. Solomonoff 85th Memorial Conference*, volume ???? of *LNAI*, pages ??–??, Melbourne, Australia, 2011. Springer.

[McC96] A. K. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, University of Rochester, 1996.

[McK09] R. B. McKenzie. *Predictably Rational? In Search of Defenses for Rational Behavior in Economics*. Springer, 2009.

[NSH11] P. Nguyen, P. Sunehag, and M. Hutter. Feature reinforcement learning in practice. In *Proc. 9th European Workshop on Reinforcement Learning (EWRL-9)*, volume 7188 of *LNAI*, pages ??–?? Springer, 2011. to appear.

[OR11] L. Orseau and M. Ring. Self-modification and mortality in artificial agents. In *Proc. 4th Conf. on Artificial General Intelligence (AGI'11)*, volume 6830 of *LNAI*, pages 1–10. Springer, Berlin, 2011.

[Ors10] L. Orseau. Optimality issues of universal greedy agents with static priors. In *Proc. 21st International Conf. on Algorithmic Learning Theory (ALT'10)*, volume 6331 of *LNAI*, pages 345–359, Canberra, 2010. Springer, Berlin.

[Pan08] S. Pankov. A computational approximation to the AIXI model. In *Proc. 1st Conference on Artificial General Intelligence*, volume 171, pages 256–267, 2008.

[Par07] M. A. Park. *Introducing Anthropology: An Integrated Approach*. McGraw-Hill, 4th edition, 2007.

[Pen94] R. Penrose. *Shadows of the Mind, A Search for the Missing Science of Consciousness*. Oxford University Press, 1994.

[PH06] J. Poland and M. Hutter. Universal learning of repeated matrix games. In *Proc. 15th Annual Machine Learning Conf. of Belgium and The Netherlands (Benelearn'06)*, pages 7–14, Ghent, 2006.

[RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, 2011.

[RN10] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition, 2010.

[RO11] M. Ring and L. Orseau. Delusion, survival, and intelligent agents. In *Proc. 4th Conf. on Artificial General Intelligence (AGI'11)*, volume 6830 of *LNAI*, pages 11–20. Springer, Berlin, 2011.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[SB09] S. Sanner and C. Boutilier. Practical solution techniques for first-order MDPs. *Artificial Intelligence*, 173(5–6):748–788, 2009.

[Sch02] J. Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Proc. 15th Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 216–228, Sydney, 2002. Springer, Berlin.

[Sch04] J. Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54(3):211–254, 2004.

[Sch07] J. Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Proc. 10th Intl. Conf. on Discovery Science (DS'07)*, volume LNAI 4755, pages 26–38, Senday, 2007. Springer.

[SDL07] A. L. Strehl, C. Diuk, and M. L. Littman. Efficient structure learning in factored-state MDPs. In *Proc. 27th AAAI Conference on Artificial Intelligence*, pages 645–650, Vancouver, BC, 2007. AAAI Press.

[Sea05] J. R. Searle. *Mind: A Brief Introduction*. Oxford University Press, USA, 2005.

[SH99] N. Suematsu and A. Hayashi. A reinforcement learning algorithm in partially observable environments using short-term memory. In *Advances in Neural Information Processing Systems 12 (NIPS'09)*, pages 1059–1065, 1999.

[SH10] P. Sunehag and M. Hutter. Consistency of feature Markov processes. In *Proc. 21st International Conf. on Algorithmic Learning Theory (ALT'10)*, volume 6331 of *LNAI*, pages 360–374, Canberra, 2010. Springer, Berlin.

[SH11a] P. Sunehag and M. Hutter. Axioms for rational reinforcement learning. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 338–352, Espoo, Finland, 2011. Springer, Berlin.

[SH11b] P. Sunehag and M. Hutter. Principles of Solomonoff induction and AIXI. In *Proc. Solomonoff 85th Memorial Conference*, volume ???? of *LNAI*, pages ??–??, Melbourne, Australia, 2011. Springer.

[SHL97] N. Suematsu, A. Hayashi, and S. Li. A Bayesian approach to model learning in non-Markovian environments. In *Proc. 14th Intl. Conf. on Machine Learning (ICML'97)*, pages 349–357, 1997.

[Ski74] B. F. Skinner. *About Behaviorism*. Random House, 1974.

[SLB09] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[SLJ+03] S. Singh, M. Littman, N. Jong, D. Pardoe, and P. Stone. Learning predictive state representations. In *Proc. 20th International Conference on Machine Learning (ICML'03)*, pages 712–719, 2003.

[SMM07] R. L. Solso, O. H. MacLin, and M. K. MacLin. *Cognitive Psychology*. Allyn & Bacon, 8th edition, 2007.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24:422–432, 1978.

[TTJ01] A. Tettamanzi, M. Tomassini, and J. Janssen. *Soft Computing: Integrating Evolutionary, Neural, and Fuzzy Systems*. Springer, 2001.

[Tur50] A. M. Turing. Computing machinery and intelligence. *Mind*, 1950.

[Tur84] R. Turner. *Logics for Artificial Intelligence*. Ellis Horwood Series in Artificial Intelligence, 1984.

[VNH+11] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011.

[VNHS10] J. Veness, K. S. Ng, M. Hutter, and D. Silver. Reinforcement learning via AIXI approximation. In *Proc. 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, pages 605–611, Atlanta, 2010. AAAI Press.

[Wal05] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, Berlin, 2005.

[WG12] P. Wang and B. Goertzel, editors. *Theoretical Foundations of Artificial General Intelligence*. Atlantis Press, Paris, 2012.

[WSH11] I. Wood, P. Sunehag, and M. Hutter. (Non-)equivalence of universal priors. In *Proc. Solomonoff 85th Memorial Conference*, volume ???? of *LNAI*, pages ??–??, Melbourne, Australia, 2011. Springer.

[WST95] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.