# Maximizing Fun by Creating Data with Easily Reducible Subjective Complexity

**Jürgen Schmidhuber**

**Abstract** The *Formal Theory of Fun and Creativity* (1990–2010) [Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans. Auton. Mental Dev. **2**(3), 230–247 (2010b)] describes principles of a curious and creative agent that never stops generating nontrivial and novel and surprising tasks and data. Two modules are needed: a data encoder and a data creator. The former encodes the growing history of sensory data as the agent is interacting with its environment; the latter executes actions shaping the history. Both learn. The encoder continually tries to encode the created data more efficiently, by discovering new regularities in it. Its *learning progress* is the *wow-effect* or *fun* or intrinsic reward of the creator, which maximizes future expected reward, being motivated to invent skills leading to interesting data that the encoder does not yet know but can easily learn with little computational effort. I have argued that this simple formal principle explains science and art and music and humor.

*Note:* This overview heavily draws on previous publications since 1990, especially Schmidhuber (2010b), parts of which are reprinted with friendly permission by IEEE.

## 1 Introduction

"All life is problem solving," wrote Popper (1999). To solve existential problems such as avoiding hunger or heat, a baby has to learn how the initially unknown environment responds to its actions. Even when there is no immediate need to satisfy

J. Schmidhuber (✉)
IDSIA, Galleria 2, 6928 Manno-Lugano, Switzerland

University of Lugano & SUPSI, Manno-Lugano, Switzerland
e-mail: juergen@idsia.ch

G. Baldassarre and M. Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*, DOI 10.1007/978-3-642-32375-1_5,

thirst or other built-in primitive drives, the baby does not run idle. Instead it actively conducts experiments: what sensory feedback do I get if I move my eyes or my fingers or my tongue just like that? Being able to predict effects of actions will later make it easier to plan control sequences leading to desirable states, such as those where heat and hunger sensors are switched off.

The growing infant quickly gets bored by things it already understands well, but also by those it does not understand at all, always searching for new effects exhibiting some yet unexplained but *easily learnable* regularity. It acquires more and more complex behaviors building on previously acquired, simpler behaviors. Eventually it might become a physicist discovering previously unknown physical laws, or an artist creating new eye-opening artworks, or a comedian coming up with novel jokes.

For a long time I have been arguing, using various wordings, that all this behavior is driven by a very simple algorithmic mechanism that uses more or less general reinforcement learning (RL) methods (Hutter 2005; Kaelbling et al. 1996; Schmidhuber 1991d, 2009e; Sutton and Barto 1998) to maximize internal *wow-effects* or *fun* or *intrinsic reward* through reductions of subjective data complexity. To make this more precise, consider two modules: a data encoder and a data creator. The former encodes the growing history of sensory data (tactile, auditory, visual, etc.) as the agent is interacting with its environment; the latter executes actions influencing and shaping the history. The encoder uses some learning algorithm to encode the data more efficiently, trying to discover new regularities that allow for saving storage space or computation time. The encoder's *progress* (the improvement through learning) is the *fun* of the creator, which uses RL to maximize future expected fun, being motivated to invent behaviors leading to interesting data that the encoder does not yet know but can easily learn.

Since 1990, agents were built that implement this idea (Schmidhuber 2010b). They may be viewed as simple artificial scientists or artists with an intrinsic desire to create experiments for building a better model of the world and of what can be done in it (Schmidhuber 1990a, 1991a,b,c, 1997c, 2002a, 2006a, 2007c, 2008, 2009a,c,d, 2010a,b, 2012; Storck et al. 1995). Crucial ingredients are the following:

1. An adaptive world model, essentially an encoder or predictor or compressor of the continually growing history of actions/events/sensory inputs, reflecting what is currently known about how the world works.
2. A learning algorithm that continually improves the encoder (detecting novel, initially surprising spatiotemporal patterns that subsequently become known patterns).
3. Intrinsic rewards measuring the encoder's improvements (first derivative of learning progress) due to the learning algorithm (thus measuring the *degree* of subjective novelty and surprise).
4. A separate reward optimizer or reinforcement learner, which translates those rewards into action sequences or behaviors expected to optimize future reward.

A simple example may help to see that it is really possible to learn from intrinsic reward signals *à la* Item **3** that one can learn even more in places never visited before. In an environment with red and blue boxes, whenever the learning agent opens a red box, it will find an easily learnable novel geometric pattern (that is, its encoder will make progress and thus generate intrinsic reward), while all blue boxes contain a generator of unpredictable, incompressible white noise. That is, all the RL creator has to learn is a simple policy: open the next unopened red box.

Ignoring issues of computation time, it is possible to devise mathematically optimal, *universal* RL methods for such systems (Schmidhuber 2006a, 2007c, 2009c,d) (2006–2009). More about this in Sect. 2. However, the practical implementations so far (Schmidhuber 1991a,b, 1997c, 2002a; Storck et al. 1995) were nonuniversal and made approximative assumptions. Among the many ways of combining algorithms for **1–4** the following variants were implemented:

**A.** Nontraditional RL for partially observable Markov decision processes (POMDPs without restrictive traditional assumptions (Schmidhuber 1991d)) based on an adaptive recurrent neural network as a predictive data encoder (Schmidhuber 1990b) is used to maximize intrinsic reward for the creator in proportion to the model's prediction errors (1990) (Schmidhuber 1990a, 1991c).

**B.** Traditional RL (Kaelbling et al. 1996; Sutton and Barto 1998) is used to maximize intrinsic reward created in proportion to *improvements* (first derivatives) of prediction error (1991) (Schmidhuber 1991a,b).

**C.** Traditional RL maximizes intrinsic reward created in proportion to relative entropies between the learning agent's priors and posteriors (1995) (Storck et al. 1995).

**D.** Nontraditional RL (Schmidhuber et al. 1997) (without restrictive Markovian assumptions) learns probabilistic, hierarchical programs and skills through zero-sum intrinsic reward games of two players, each trying to out-predict or surprise the other, taking into account the computational costs of learning, and learning *when* to learn and *what* to learn (1997–2002) (Schmidhuber 1997c, 2002a).

**B–D.** (1991–2002) also showed experimentally how intrinsic rewards can substantially accelerate goal-directed learning and *external* reward intake.

**Outline.** Section 2 will summarize the formal theory of creativity in a nutshell, laying out a mathematically rigorous but not necessarily practical framework. Section 3 will then discuss previous concrete implementations of the nonoptimal but currently still more practical variants **A–D** mentioned above and their limitations. Section 4 will discuss relations to work by others and show how the theory greatly extends the traditional field of active learning and how it formalizes and extends previous informal ideas of developmental psychology and aesthetics theory. Section 5 will offer a natural typology of computational intrinsic motivation (IM), and Sect. 6 will briefly explain how the theory is indeed sufficiently general to explain all kinds of creative behavior, from the discovery of new physical laws through active design of experiments to the invention of jokes and works of art.

## 2 Formal Details of the Theory of Creativity

The theory formulates essential principles behind numerous intrinsically motivated *creative* behaviors of biological or artificial agents embedded in a possibly unknown environment. The corresponding algorithmic framework uses general RL (Sect. 2.6; Hutter 2005; Schmidhuber 2009e) to maximize not only external reward for achieving goals such as the satisfaction of hunger and thirst but also *intrinsic* reward or *wow-effect* reward for learning a better data encoder or model, by creating/discovering/learning novel patterns in the growing history of actions and sensory inputs, where the theory formally specifies what exactly is a *pattern*, what exactly is *novel* or surprising, and what exactly it means to incrementally *learn* novel skills, leading to more novel patterns.

### *2.1 The Creative Agent and Its Improving Data Encoder*

Let us consider a learning agent whose single life consists of discrete cycles or time steps $t = 1, 2, \ldots, T$. Its complete lifetime $T$ may or may not be known in advance. In what follows, the value of any time-varying variable $Q$ at time $t$ ($1 \leq t \leq T$) will be denoted by $Q(t)$, the ordered sequence of values $Q(1), \ldots, Q(t)$ by $Q(\leq t)$, and the (possibly empty) sequence $Q(1), \ldots, Q(t-1)$ by $Q(<t)$. At any given $t$, the agent receives a real-valued input $x(t)$ from the environment and executes a real-valued action $y(t)$ which may affect future inputs. At times $t < T$, its goal is to maximize future success or *utility*

$$u(t) = E_\mu \left[ \sum_{\tau=t+1}^{T} r(\tau) \;\middle|\; h(\leq t) \right], \tag{1}$$

where the reward $r(t)$ is a special real-valued input (vector) at time $t$, $h(t)$ the ordered triple $[x(t), y(t), r(t)]$ (hence $h(\leq t)$ is the known history up to $t$), and $E_\mu(\cdot \mid \cdot)$ denotes the conditional expectation operator with respect to some possibly unknown distribution $\mu$ from a set $\mathcal{M}$ of possible distributions. Here $\mathcal{M}$ reflects whatever is known about the possibly probabilistic reactions of the environment. As a very general example, $\mathcal{M}$ may contain all computable distributions (Hutter 2005; Li and Vitányi 1997; Solomonoff 1978). This essentially includes all environments one could write scientific papers about. There is just one life, no need for predefined repeatable trials, no restriction to Markovian interfaces between sensors and environment (Schmidhuber 1991d). (Note that traditional Markovian RL (Sutton and Barto 1998) typically is too limited as it assumes the current input tells the agents everything it needs to know and does not work in realistic scenarios where robots have to learn to memorize previous relevant inputs in the form of appropriate internal representations.) The utility function implicitly takes

into account the expected remaining lifespan $E_\mu(T \mid h(\leq t))$ and thus the possibility to extend the lifespan through appropriate actions (Schmidhuber 2005a, 2009e). Note that mathematical analysis is *not* simplified by discounting future rewards like in traditional RL theory (Sutton and Barto 1998)—one should avoid such distortions of real rewards whenever possible.

To maximize $u(t)$, the agent may profit from an adaptive, predictive *model* $p$ of the consequences of its possible interactions with the environment. At any time $t$ ($1 \leq t < T$), the model $p(t)$ will depend on the observed history so far, $h(\leq t)$. It may be viewed as the current explanation or description of $h(\leq t)$ and may help to predict future events, including rewards. Let $C(p, h)$ denote some given model $p$'s quality or performance evaluated on a given history $h$. Natural quality measures taking into account storage space and computation time will be discussed in Sect. 2.2.

To encourage the agent to actively create data leading to easily learnable improvements of $p$ (Schmidhuber 1991b,c, 2002a, 2006a, 2007c, 2009a,c,d, 2010a,b; Storck et al. 1995), the reward signal $r(t)$ is simply split into two scalar real-valued components: $r(t) = g(r_{\text{ext}}(t), r_{\text{int}}(t))$, where $g$ maps pairs of real values to real values, for example, $g(a, b) = a + b$. Here $r_{\text{ext}}(t)$ denotes traditional *external* reward provided by the environment, such as negative reward in response to bumping against a wall or positive reward in response to reaching some teacher-given goal state. The formal theory of creativity, however, is especially interested in $r_{\text{int}}(t)$, the internal or *intrinsic* reward, which is provided whenever the data encoder's quality improves—for *purely creative* agents $r_{\text{ext}}(t) = 0$ for all valid $t$:

> The current *fun* $r_{\text{int}}(t)$ of the action selector is measured by the *improvements* of the data encoder $p$ at time $t$.

Formally, the intrinsic reward in response to the encoder's progress (due to some application-dependent model improvement algorithm) between times $t$ and $t + 1$ is

$$r_{\text{int}}(t + 1) = f\,[C(p(t), h(\leq t + 1)), C(p(t + 1), h(\leq t + 1))], \tag{2}$$

where $f$ maps pairs of real values to real values. Various alternative progress measures are possible; most obvious is $f(a, b) = a - b$. This corresponds to a discrete time version of maximizing the first derivative of the encoder's quality. *Note that both the old and the new encoder have to be tested on the same data, namely, the history so far.* So progress between times $t$ and $t + 1$ is defined based on two encoders of $h(<= t + 1)$, where the old one is trained only on $h(<= t)$ and the new one also gets to see $h(t <= t + 1)$. This is like $p(t)$ predicting data of time $t + 1$, then observing it, then learning something, and then becoming a measurably better encoder $p(t + 1)$.

The above description of the agent's motivation conceptually separates the goal (finding or creating data that can be modeled better or faster than before) from the means of achieving the goal. Let the creator's RL mechanism figure out how to translate such rewards into action sequences that allow the given encoder improvement algorithm to find and exploit previously unknown types of

regularities. It is the task of the RL algorithm to trade off long-term versus short-term intrinsic rewards of this kind, taking into account all costs of action sequences (Schmidhuber 1991b,c, 2002a, 2006a, 2007c, 2009a,c,d, 2010a,b; Storck et al. 1995). The universal RL methods of Sect. 2.6 as well as RNN-based RL (Sect. 3.1) and SSA-based RL (Sect. 3.4) can in principle learn useful internal states containing memories of relevant previous events; less powerful RL methods (Sects. 3.2 and 3.3) cannot.

## 2.2 How to Measure Encoder Quality Under Time Constraints

In theory $C(p, h(\leq t))$ should take the entire history of actions and perceptions into account (Schmidhuber 2006a), like the following performance measure $C_{xry}$:

$$C_{xry}(p, h(\leq t)) = \sum_{\tau=1}^{t} || \, pred(p, x(\tau)) - x(\tau) \, ||^2 + || \, pred(p, r(\tau)) - r(\tau) \, ||^2 + || \, pred(p, y(\tau)) - y(\tau) \, ||^2 \quad (3)$$

where $pred(p, q)$ is $p$'s prediction of event $q$ from earlier parts of the history (Schmidhuber 2006a).

$C_{xry}$ ignores the danger of overfitting through a $p$ that just stores the entire history without compactly representing its regularities, if any. The principle of minimum description length (MDL) (Kolmogorov 1965; Li and Vitányi 1997; Rissanen 1978; Solomonoff 1978; Wallace and Boulton 1968; Wallace and Freeman 1987), however, also takes into account the description size of $p$, viewing $p$ as a compressor program of the data $h(\leq t)$. This program $p$ should be able to deal with any prefix of the growing history, computing an output starting with $h(\leq t)$ for any time $t$ $(1 \leq t < T)$. (A program that wants to halt after $t$ steps can easily be fixed/augmented by the trivial method that simply stores any raw additional data coming in after the halt.)

$C_l(p, h(\leq t))$ denotes $p$'s compression performance on $h(\leq t)$: the number of bits needed to specify both the encoder and the deviations of the sensory history from its predictions, in the sense of loss-free compression. The smaller the $C_l$, the more regularity and compressibility and predictability and lawfulness in the observations so far.

For example, suppose $p$ uses a small encoder that correctly predicts many $x(\tau)$ for $1 \leq \tau \leq t$. This can be used to encode $x(\leq t)$ compactly: given the predictor, only the wrongly predicted $x(\tau)$ plus information about the corresponding time steps $\tau$ are necessary to reconstruct input history $x(\leq t)$, for example (Schmidhuber 1992). Similarly, a predictor that learns a probability distribution on the possible next events, given previous events, can be used to efficiently encode observations

with high (respectively low) predicted probability by few (respectively many) bits (Sect. 3.3; (Huffman 1952; Schmidhuber and Heil 1996)), thus achieving a compressed history representation.

Alternatively, $p$ could also make use of a 3D world model or simulation. The corresponding MDL-based quality measure $C_{3D}(p, h(\leq t))$ is the number of bits needed to specify all polygons and surface textures in the 3D simulation plus the number of bits needed to encode deviations of $h(\leq t)$ from the predictions of the simulation. Improving the 3D model by adding or removing polygons may reduce the total number of bits required.

The ultimate limit for $C_l(p, h(\leq t))$ would be $K^*(h(\leq t))$, a variant of the Kolmogorov complexity of $h(\leq t)$, namely, the length of the shortest program (for the given hardware) that computes an output starting with $h(\leq t)$ (Kolmogorov 1965; Li and Vitányi 1997; Schmidhuber 2002b; Solomonoff 1978). Here there is no need to worry about the fact that $K^*(h(\leq t))$ in general cannot be computed exactly, only approximated from above (indeed, for most practical predictors, the approximation will be crude). This just means that some patterns will be hard to detect by the limited encoder of choice, that is, the reward maximizer will get discouraged from spending too much effort on creating those patterns.

$C_l(p, h(\leq t))$ does not take into account the time $\tau(p, h(\leq t))$ spent by $p$ on computing $h(\leq t)$. A runtime-dependent performance measure inspired by concepts of optimal universal search (Levin 1973; Schmidhuber 2002c, 2004, 2006a, 2009c, 2010a) is

$$C_{l\tau}(p, h(\leq t)) = C_l(p, h(\leq t)) + \log\ \tau(p, h(\leq t)). \tag{4}$$

Here compression by one bit is worth as much as runtime reduction by a factor of $\frac{1}{2}$. From an asymptotic optimality-oriented point of view, this is one of the best ways of trading off storage and computation time (Levin 1973; Schmidhuber 2002c, 2004).

In practical applications (Sect. 3), the encoder of the continually growing data typically will have to calculate its output online, that is, it will be able to use only a constant number of computational instructions per second to predict/compress new data. The goal of the possibly much slower learning algorithm must then be to improve the encoder such that it keeps operating online within those time limits, while encoding better than before. The costs of computing $C_{xry}(p, h(\leq t))$ and $C_l(p, h(\leq t))$ and similar performance measures are linear in $t$, assuming $p$ consumes equal amounts of computation time for each single prediction. Therefore, online evaluations of learning progress on the full history so far generally cannot take place as frequently as the continually ongoing online predictions.

At least some of the learning and its progress evaluations may take place during occasional "sleep" phases (Schmidhuber 2006a, 2009c). But practical implementations so far have looked only at parts of the history for efficiency reasons: the systems described in Sects. 3 and 3.1–3.4 (Schmidhuber 1991b,c, 2002a; Storck et al. 1995) used online settings (one prediction per time step and constant computational effort per prediction), nonuniversal adaptive encoders, and approximative evaluations of learning progress, each consuming only constant time despite the continual growth of the history.

### 2.3 Optimal Predictors Versus Optimal Compressors

For the theoretically inclined, there is a deep connection between optimal prediction and optimal compression. Consider Solomonoff's theoretically optimal, universal way of predicting the future (Hutter 2005; Li and Vitányi 1997; Solomonoff 1964, 1978). Given an observation sequence $q(\leq t)$, the Bayes formula is used to predict the probability of the next possible $q(t+1)$. The only assumption is that there exists a computer program that can take any $q(\leq t)$ as an input and compute its a priori probability according to the $\mu$ prior. (This assumption is extremely general, essentially including all environments one can write scientific papers about, as mentioned above.) In general this program is unknown; hence, a mixture prior is used instead to predict:

$$\xi(q(\leq t)) = \sum_i w_i \mu_i(q(\leq t)), \tag{5}$$

a weighted sum of *all* distributions $\mu_i \in \mathcal{M}$, $i = 1, 2, \ldots$, where the sum of the constant positive weights satisfies $\sum_i w_i \leq 1$. This is indeed the best one can possibly do, in a very general sense (Hutter 2005; Solomonoff 1978). The drawback of the scheme is its incomputability, since $\mathcal{M}$ contains infinitely many distributions. One may increase the theoretical power of the scheme by augmenting $\mathcal{M}$ by certain non-enumerable but limit-computable distributions (Schmidhuber 2002b) or restrict it such that it becomes computable, for example, by assuming the world is computed by some unknown but deterministic computer program sampled from the speed prior (Schmidhuber 2002c) which assigns low probability to environments that are hard to compute by any method.

Remarkably, under very general conditions, both universal inductive inference (Li and Vitányi 1997; Solomonoff 1964, 1978) and the compression-oriented MDL approach (Kolmogorov 1965; Li and Vitányi 1997; Rissanen 1978; Wallace and Boulton 1968; Wallace and Freeman 1987) converge to the correct predictions in the limit (Poland and Hutter 2005). It should be mentioned, however, that the former converges faster.

### 2.4 Discrete Asynchronous Framework for Maximizing Fun

Let $p(t)$ denote the agent's current encoder program at time $t$, $s(t)$ its current creator, and **DO:**

**Creator:** At any time $t$ $(1 \leq t < T)$ do:

1. Let $s(t)$ use (parts of) history $h(\leq t)$ to select and execute $y(t+1)$.
2. Observe $x(t+1)$.
3. Check if there is nonzero intrinsic reward $r_{\text{int}}(t+1)$ provided by the asynchronously running improvement algorithm of the encoder (see below). If not, set $r_{\text{int}}(t+1) = 0$.

4. Let the creator's RL algorithm use $h(\leq t+1)$ including $r_{\text{int}}(t+1)$ (and possibly also the latest available compressed version of the observed data—see below) to obtain a new creator $s(t+1)$, in line with objective (1). Note that some actions may actually trigger learning algorithms that compute changes of the encoder and the creator's policy, such as in Sect. 3.4 (Schmidhuber 2002a). That is, the computational cost of learning can be taken into account by the reward optimizer, and the decision when and what to learn can be learned as well (Schmidhuber 2002a).

**Encoder:** Set $p_{\text{new}}$ equal to the initial data encoder. Starting at time 1, repeat forever until interrupted by death at time $T$:

1. Set $p_{\text{old}} = p_{\text{new}}$, get current time step $t$, and set $h_{\text{old}} = h(\leq t)$.
2. Evaluate $p_{\text{old}}$ on $h_{\text{old}}$ to obtain performance measure $C(p_{\text{old}}, h_{\text{old}})$. This may take many time steps.
3. Let some (possibly application-dependent) encoder improvement algorithm (such as a learning algorithm for an adaptive neural network predictor, possibly triggered by a creator action) use $h_{\text{old}}$ to obtain a hopefully better encoder $p_{\text{new}}$ (such as a neural net with the same size and the same constant computational effort per prediction but with improved predictive power and therefore improved compression performance (Schmidhuber and Heil 1996)). Although this may take many time steps (and could be partially performed off-line during "sleep" (Schmidhuber 2006a, 2009c)), $p_{\text{new}}$ may not be optimal due to limitations of the learning algorithm, for example, local maxima. (To inform the creator about beginnings of encoder evaluation processes, etc., augment its input by unique representations of such events.)
4. Evaluate $p_{\text{new}}$ on $h_{\text{old}}$ to obtain $C(p_{\text{new}}, h_{\text{old}})$. This may take many time steps.
5. Get current time step $\tau$ and generate fun

$$r_{\text{int}}(\tau) = f[C(p_{\text{old}}, h_{\text{old}}), C(p_{\text{new}}, h_{\text{old}})], \tag{6}$$

for example, $f(a,b) = a - b$. [Here the $\tau$ replaces the $t+1$ of Eq. (2).]

This asynchronuous scheme (Schmidhuber 2006a, 2007c, 2009c) may cause long temporal delays between creator actions and corresponding rewards or fun events and may impose a heavy burden on the creator's RL algorithm whose task is to assign credit to past actions. Nevertheless, Sect. 2.6 will discuss RL algorithms for this purpose which are theoretically optimal in various senses (Schmidhuber 2006a, 2007c, 2009c,d).

## *2.5 Continuous Time Formulation*

In continuous time, $O(t)$ denotes the state of subjective observer $O$ at time $t$. The subjective computational complexity $Complexity(D, O(t))$ of a sequence of

observations and/or actions is a measure of the effort required to encode/decode $D$, given $O(t)$'s current limited prior knowledge and limited encoding method. The time-dependent and observer-dependent subjective $Fun(D, O(t))$ is

$$Fun(D, O(t)) \sim \frac{\partial Complexity(D, O(t))}{\partial t}, \tag{7}$$

the *first derivative* of subjective complexity or simplicity: as $O$ improves its encoder, formerly apparently random data parts become subjectively more regular/elegant/beautiful, requiring fewer bits (or less time) for their encoding.

There are at least two ways of having fun: execute a learning algorithm that improves the compression of the already known data (in online settings, without increasing computational needs of the encoder) or execute actions that generate more data, then learn to better encode this new data.

## 2.6 Optimal Creativity, Given the Encoder's Limitations

The previous sections discussed how to measure encoder improvements and how to translate them into intrinsic reward signals, but did not say much about the RL method used to maximize expected future reward. The chosen encoder class typically will have certain computational limitations. In the absence of any external rewards, one may define *optimal pure curiosity behavior* relative to these limitations: at discrete time step $t$, this behavior would select the action that maximizes

$$u(t) = E_\mu \left[ \sum_{\tau=t+1}^{T} r_{\text{int}}(\tau) \;\middle|\; h(\leq t) \right]. \tag{8}$$

Since the true, world-governing probability distribution $\mu$ is unknown, the resulting task of the creator's RL algorithm may be a formidable one. As the system is revisiting previously incompressible parts of the environment, some of those will tend to become more subjectively compressible, and the corresponding curiosity rewards will decrease over time. A good RL algorithm must somehow detect and then *predict* this decrease and act accordingly. Traditional RL algorithms (Kaelbling et al. 1996), however, do not provide any theoretical guarantee of optimality for such situations.

Is there a best possible RL algorithm that comes as close as any other computable one to maximizing objective (8)? Indeed, there is. Its drawback, however, is that it is not computable in finite time. Nevertheless, it serves as a reference point for defining what is achievable at best, that is, what is *optimal* creativity.

How does it work? Optimal inductive inference as defined in Sect. 2.3 can be extended by formally including the effects of executed actions, to define an optimal action selector maximizing future expected reward. At any time $t$, Hutter's

theoretically optimal (yet uncomputable) RL algorithm AIXI (Hutter 2005) uses such an extended version of Solomonoff's scheme to select those action sequences that promise maximal future reward up to some horizon $T$ (e.g., twice the lifetime so far), given the current data $h(\leq t)$. That is, in cycle $t + 1$, AIXI selects as its next action the first action of an action sequence maximizing $\xi$-predicted reward up to the given horizon, appropriately generalizing Eq. (5). AIXI uses observations optimally (Hutter 2005): the Bayes-optimal policy $p^{\xi}$ based on the mixture $\xi$ is self-optimizing in the sense that its average utility value converges asymptotically for all $\mu \in \mathcal{M}$ to the optimal value achieved by the Bayes-optimal policy $p^{\mu}$ which knows $\mu$ in advance. The necessary and sufficient condition is that $\mathcal{M}$ admits self-optimizing policies. The policy $p^{\xi}$ is also Pareto optimal in the sense that there is no other policy yielding higher or equal value in *all* environments $\nu \in \mathcal{M}$ and a strictly higher value in at least one (Hutter 2005).

AIXI as above needs unlimited computation time. Its computable variant AIXI*(t,l)* (Hutter 2005) has asymptotically optimal runtime but may suffer from a huge constant slowdown. To take the consumed computation time into account in a general, optimal way, one may use the recent Gödel machine s (Schmidhuber 2005a,b, 2006b, 2009e) instead. They represent the first class of mathematically rigorous, fully self-referential, self-improving, general, optimally efficient problem solvers and are applicable to the problem embodied by objective (8).

The initial software $\mathcal{S}$ of such a Gödel machine contains an initial problem solver, for example, some typically suboptimal RL method (Kaelbling et al. 1996). It also contains an asymptotically optimal initial proof searcher, typically based on an online variant of Levin's *universal search* (Levin 1973), which is used to run and test *proof techniques*. Proof techniques are programs written in a universal language implemented on the Gödel machine within $\mathcal{S}$. They are in principle able to compute proofs concerning the system's own future performance, based on an axiomatic system $\mathcal{A}$ encoded in $\mathcal{S}$. $\mathcal{A}$ describes the formal *utility* function, in the present case Eq. (8), the hardware properties, axioms of arithmetic and probability theory and data manipulation, etc., and $\mathcal{S}$ itself, which is possible without introducing circularity (Schmidhuber 2009e).

Inspired by Kurt Gödel's celebrated self-referential formulas (1931), the Gödel machine rewrites any part of its own code (including the proof searcher) through a self-generated executable program as soon as its *universal search* variant has found a proof that the rewrite is *useful* according to objective (8). According to the Global Optimality Theorem (Schmidhuber 2005a,b, 2006b, 2009e), such a self-rewrite is globally optimal—no local maxima possible!—since the self-referential code first had to prove that it is not useful to continue the search for alternative self-rewrites.

If there is no provably useful optimal way of rewriting $\mathcal{S}$ at all, then humans will not find one either. But if there is one, then $\mathcal{S}$ itself can find and exploit it. Unlike the previous *non*-self-referential methods based on hardwired proof searchers (Hutter 2005), Gödel machine s not only boast an optimal *order* of complexity but can optimally reduce (through self-changes) any slowdowns hidden by the $O()$-notation, provided the utility of such speedups is provable (Schmidhuber 2006c, 2007a,b).

**Limitations of the "Universal" Approaches.** The methods above are optimal in various ways, some of them not only computable but even optimally time efficient in the asymptotic limit. Nevertheless, they leave open an essential remaining practical question: If the agent can execute only a fixed number of computational instructions per unit time interval (say, 10 trillion elementary operations per second), what is the best way of using them to get as close as possible to the theoretical limits of universal AIs? Especially when external rewards are very rare, as is the case in many realistic environments? As long as there is no good answer to this question, one has to resort to approximations and heuristics when it comes to practical applications. The next section reviews what has been achieved so far along these lines, discussing our implementations of curiosity-driven creative agents from the 1990s; quite a few aspects of these concrete systems are still of relevance today.

## 3 Previous Implementations of Curious/Creative Agents: Pros and Cons

The above mathematically rigorous framework for optimal curiosity and creativity (2006–) was established *after* first approximations thereof were implemented (1991, 1995, 1997–2002). Sections 3.1–3.4 will discuss advantages and limitations of online learning systems described in the original publications on artificial intrinsic motivation (Schmidhuber 1991b,c, 1997c; Storck et al. 1995) which already can be viewed as example implementations of a compression progress drive or prediction progress drive that encourages the discovery or creation of novel, surprising patterns. Some elements of this earlier work are believed to remain essential for creating systems that are both theoretically sound and *practical*.

### *3.1 Intrinsic Reward for Coding Error (1990)*

Early work (Schmidhuber 1990a, 1991c) describes a data encoder based on an adaptive world model implemented as a recurrent neural network (RNN) (in principle, a rather powerful computational device, even by today's machine learning standards), predicting sensory inputs including reward signals from the entire previous history of actions and inputs. A second RNN (the creator) uses the world model and gradient descent to search for a control policy or program maximizing the sum of future expected rewards according to the model. Some of the rewards are intrinsic curiosity rewards, which are proportional to the predictor's errors. So the same mechanism that is used for normal goal-directed learning is used for implementing creativity and curiosity and boredom—there is no need for a separate system aiming at improving the encoder.

This first description of a general, curious, world-exploring RL agent implicitly and optimistically assumes that the encoder will indeed improve by motivating the

creator/controller to go to places where the prediction error is high. One drawback of the prediction error-based approach is that it encourages the creator to focus its search on those parts of the environment where there will always be high prediction errors due to noise or randomness or due to computational limitations of the encoder. This may *prevent* learning progress instead of promoting it and motivates the next subsection, whose basic ideas could be combined with the RL method of (Schmidhuber 1990a, 1991c), but this has not been done yet.

Another potential drawback is the nature of the particular RNN-based RL method. Although the latter has the potential to learn internal memories of previous relevant sensory inputs, and thus can deal with POMDPs as it is not limited to Markovian interfaces between agent and environment (Schmidhuber 1991d), like all gradient-based methods, it may suffer from local minima, as well as from potential problems of online learning, since gradients for the recurrent RL creator are computed with the help of the dynamically changing, online learning recurrent predictive encoder. Apart from this limitation, the RNN of back then were less powerful than today's LSTM RNN (Hochreiter and Schmidhuber 1997; Schmidhuber et al. 2011), which yielded state of the art performance in challenging applications such as connected handwriting recognition (Graves et al. 2009), and should be used instead.

## *3.2 Intrinsic Reward for Encoder Improvements (1991)*

Follow-up work (Schmidhuber 1991a,b) points out that one should not focus on the errors of the encoder, but on its improvements. The basic principle can be formulated as follows: *Learn a mapping from actions (or action sequences) to the expectation of future performance improvement of the encoder. Encourage action sequences where this expectation is high.*

Two implementations were described: The first models the reliability of the predictions of the encoder/predictor by a separate, so-called confidence network. At any given time, reinforcement for the model-building control system is created in proportion to the current *change* or first derivative of the reliability of the adaptive predictor. The "curiosity goal" of the creator (it might have additional "pre-wired" external goals) is to maximize the expectation of the cumulative sum of future positive or negative changes in prediction reliability.

The second implementation replaces the confidence network by a network $H$ which at every time step tries to predict the current *change* or first derivative of the model network's output (caused by its learning algorithm). $H$ will learn approximations of the expectations of the changes (first derivatives) of the encoder's responses to given inputs. The *absolute value* of $H$'s output is taken as the intrinsic reward.

While the neural predictor of the implementations is computationally less powerful than the recurrent one of Sect. 3.1 (Schmidhuber 1991c), there is a novelty, namely, an explicit (neural) adaptive model of the predictor's improvements,

measured in terms of mean squared error (MSE). This model essentially learns to predict the encoder's changes (the prediction derivatives). For example, although noise is unpredictable and leads to wildly varying target signals for the predictor, in the long run, these signals do not change the adaptive encoder's parameters much, and the predictor of predictor changes is able to learn this. A variant of the standard RL algorithm Q-learning (Sutton and Barto 1998) is fed with curiosity reward signals proportional to the expected long-term predictor changes; thus, the agent is intrinsically motivated to make novel patterns within the given limitations. In fact, one may say that the system tries to maximize an approximation of the (discounted) sum of the expected first derivatives of the data's subjective predictability, thus also maximizing an approximation of the (discounted) sum of the expected changes of the data's subjective compressibility (the surprise or novelty).

Both variants avoid the theoretically desirable but impractical regular evaluations of the encoder on the entire history so far, as discussed in Sect. 2.2. Instead they monitor the recent effects of learning on the learning mechanism (a neural network in this case). Experiments illustrate the advantages of this type of directed, curious exploration over traditional random exploration.

One RL method-specific drawback is given by the limitations of standard Markovian RL (Schmidhuber 1991d), which assumes the current input tells the agent everything it needs to know and does not work well in realistic scenarios where it has to learn to memorize previous relevant inputs to select optimal actions. For general robot scenarios, more powerful RL methods are necessary, such as those mentioned in Sect. 3.1 and other parts of this chapter.

Any RL algorithm has to deal with the fact that intrinsic rewards vanish where the encoder becomes perfect. In the simple toy world (Schmidhuber 1991a,b),this is not a problem, since the creator continually updates its Q-values based on recent experience. But since the learning rate is chosen heuristically (as usual in RL applications), this approach lacks the theoretical justification of the general framework of Sect. 2.

For probabilistic worlds, there are prediction error measures that are more principled than MSE. This motivates research described next.

### *3.3 Fun Depending on the Relative Entropy Between Encoder's Prior and Posterior (1995)*

Follow-up work (1995) describes an information theory-oriented variant of the approach in nondeterministic worlds (Storck et al. 1995). Here the intrinsic reward is proportional to the encoder's surprise/information gain (Fedorov 1972), measured as the Kullback–Leibler distance (Kullback 1959) between a learning predictor's subjective probability distributions before and after new observations—the relative entropy between its prior and posterior, essentially another measure of learning progress which does not take into account computation time. Again experiments show the advantages of this type of curious exploration over conventional random exploration.

Since this implementation also uses a traditional RL method (Sutton and Barto 1998) instead of a more general one, the discussion of RL method-specific drawbacks in previous subsections remains valid here as well.

Note the connection to Sect. 2: the concepts of Huffman coding (Huffman 1952) and relative entropy between prior and posterior immediately translate into a measure of learning progress reflecting the number of saved bits—a measure of improved data compression.

Note also, however, that the naive probabilistic approach to data compression is unable to discover more general types of *algorithmic* compressibility (Li and Vitányi 1997) as discussed in Sect. 2. For example, the decimal expansion of $\pi$ looks random and incompressible but is not: there is a very short algorithm computing all of $\pi$, yet any finite sequence of digits will occur in $\pi$'s expansion as frequently as expected if $\pi$ were truly random, that is, no simple statistical learner will outperform random guessing at predicting the next digit from a limited time window of previous digits. More general *program* search techniques are necessary to extract the underlying algorithmic regularity. This motivates the universal approach discussed in Sect. 2 but also the research on a more general practical implementation described next.

### *3.4 Learning Better Encodings and Skills Through Zero-Sum Intrinsic Reward Games (1997–2002)*

The universal variants of the principle of novel pattern creation of Sect. 2 focused on theoretically optimal ways of measuring encoder progress/surprise/novelty, as well as mathematically optimal ways of selecting action sequences or experiments within the framework of artificial curiosity and creativity (Schmidhuber 2006a, 2007c, 2009c,d, 2010b). These variants take the entire lifelong history of actions and observations into account and make minimal assumptions about the nature of the environment, such as the following: the (unknown) probabilities of possible event histories are at least enumerable. The resulting systems exhibit "mathematically optimal curiosity and creativity" and provide a yardstick against which all less universal intrinsically motivated systems can be measured. However, most of them ignore important issues of time constraints in online settings. For example, in practical applications, one cannot frequently measure encoder improvements by testing encoder performance on the entire history so far. The costs of learning and testing have to be taken into account. This insight drove the research discussed next.

To address the computational costs of learning and the costs of measuring learning progress, computationally powerful encoders and creators (Schmidhuber 1997c, 2002a) were implemented as two very general, coevolving, symmetric, opposing modules called the *right brain* and the *left brain*, both able to construct self-modifying probabilistic programs written in a universal programming language (1997–2002). An internal storage for temporary computational results of

the programs is viewed as part of the changing environment. Each module can suggest experiments in the form of probabilistic algorithms to be executed and make predictions about their effects, *betting intrinsic reward* on their outcomes. The opposing module may accept such a bet in a zero-sum game by making a contrary prediction or reject it. In case of acceptance, the winner is determined by executing the algorithmic experiment and checking its outcome; the intrinsic reward for wow-effects eventually gets transferred from the surprised loser to the confirmed winner. Both modules try to maximize reward using a rather general RL algorithm [the so-called success-story algorithm SSA (Schmidhuber et al. 1997)] designed for complex stochastic policies—alternative RL algorithms could be plugged in as well. Thus, both modules are motivated to discover *truly novel* algorithmic patterns/compressibility, where the subjective baseline for novelty is given by what the opponent already knows about the (external or internal) world's repetitive patterns. Since the execution of any computational or physical action costs something (as it will reduce the cumulative reward per time ratio), both modules are motivated to focus on those parts of the dynamic world that currently make learning progress *easy*, to minimize the costs of identifying promising experiments and executing them. The system learns a partly hierarchical structure of more and more complex skills or programs necessary to solve the growing sequence of self-generated tasks, reusing previously acquired simpler skills where this is beneficial. Experimental studies exhibit several sequential stages of emergent developmental sequences, with and without external reward.

Many ingredients of this system may be just what one needs to build *practical yet sound* curious and creative systems that never stop expanding their knowledge about what can be done in a given world, although future reimplementations should probably use alternative reward optimizers that are more general and powerful than SSA (Schmidhuber et al. 1997), such as variants of the Optimal Ordered Problem Solver (Schmidhuber 2004).

### *3.5 Recent Implementations*

More recently implemented variants deal with applications to vision-based reinforcement learning/evolutionary search (Cuccu et al. 2011; Luciw et al. 2011), active learning of currently easily learnable functions (Ngo et al. 2011), black-box optimization (Schaul et al. 2011b), and detection of "interesting" sequences of Wikipedia articles (Schaul et al. 2011a).

### *3.6 Improving Real Reward Intake (1991–)*

References above demonstrated through several experiments that the presence of intrinsic reward or curiosity reward can actually speed up the collection of *external*

reward. However, the previous papers also pointed out that it is always possible to design environments where the bias toward regularities introduced through artificial curiosity can lead to worse performance—curiosity can indeed kill the cat.

## 4 Relation to Work by Others

### *4.1 Beyond Traditional Information Theory and Active Learning*

How does the notion of surprise in the theory of creativity differ from the notion of surprise in traditional information theory? Consider two extreme examples of uninteresting, unsurprising, boring data: a vision-based agent that always stays in the dark will experience an extremely compressible, soon totally predictable history of unchanging visual inputs. In front of a screen full of white noise conveying a lot of information and "novelty" and "surprise" in the traditional sense of Boltzmann and Shannon (Shannon 1948), however, it will experience highly unpredictable and fundamentally incompressible data. According to the theory of creativity, in both cases, the data is not *surprising* but *boring* (Schmidhuber 2002a, 2007c) as it does not allow for further encoding progress—there is no novel pattern. Therefore, the traditional notion of surprise is rejected. Neither the arbitrary nor the fully predictable is *truly* novel or surprising. Only data with still *unknown* algorithmic regularities are (Schmidhuber 1991b,c, 2002a, 2006a, 2007c, 2009c,d; Storck et al. 1995), for example, a previously unknown song containing a subjectively novel harmonic pattern. That is why one really has to measure the *progress of the learning encoder* to compute the degree of surprise. (Compare Sect. 4.4.2 for a related discussion on what is aesthetically pleasing.)

How does the theory generalize the related traditional field of *active learning*, for example (Fedorov 1972)? To optimize a function may require expensive data evaluations. Original active learning is limited to supervised classification tasks, for example (Balcan et al. 2009; Cohn 1994; Fedorov 1972; Hwang et al. 1991; MacKay 1992; Plutowski et al. 1994; Seung et al. 1992), asking which data points to evaluate next to maximize information gain, typically (but not necessarily) using one step look-ahead, assuming all data point evaluations are equally costly. The objective (improve classification error) is given externally; there is no explicit intrinsic reward in the sense discussed in this chapter. The more general framework of creativity theory also takes formally into account:

1. Reinforcement learning agents embedded in an environment where there may be arbitrary delays between experimental actions and corresponding information gains, for example (Schmidhuber 1991b; Storck et al. 1995)
2. The highly environment-dependent costs of obtaining or creating not just individual data points but data *sequences* of *a priori* unknown size

3. Arbitrary algorithmic or statistical dependencies in sequences of actions and sensory inputs, for example (Schmidhuber 2002a, 2006a)
4. The computational cost of learning new skills, for example (Schmidhuber 2002a)

While others recently have started to study active RL as well, for example, Brafman and Tennenholtz [R-max algorithm (Brafman and Tennenholtz 2002)] and Li et al. [KWIK-framework (Li et al. 2008)] (Strehl et al. 2010), our more general systems measure and maximize *algorithmic* (Kolmogorov 1965; Li and Vitányi 1997; Schmidhuber 2002b; Solomonoff 1978) novelty (learnable but previously unknown compressibility or predictability) of self-generated spatiotemporal patterns in the history of data and actions (Schmidhuber 2006a, 2007c, 2009c,d).

## *4.2 Relation to Handcrafted Interestingness*

Lenat's discovery system EURISKO (Lenat 1983) has a preprogrammed interestingness measure which was observed to become more and more inappropriate ("stagnation" problem) as EURISKO creates new concepts from old ones with the help of human intervention. Unsupervised systems based on creativity theory, however, continually redefine what is interesting based on what is currently easy to learn, in addition to what is already known.

## *4.3 Related Implementations Since 2005*

In 2005, Baldi and Itti demonstrated experimentally that the method of 1995 (Sect. 3.3, Storck et al. 1995) explains certain patterns of human visual attention better than certain previous approaches (Itti and Baldi 2005).

Klyubin et al.'s seemingly related approach to intrinsic motivation (Klyubin et al. 2005) of 2005 tries to maximize *empowerment* by maximizing the information an agent could potentially "inject" into its future sensory inputs via a sequence of actions. The authors assume a good world model is already given or at least learned before *empowerment* is measured (D. Polani, personal communication, 2010). For example, using one step look-ahead in a deterministic and well-modeled world, their agent will prefer states where the execution of alternative actions will make a lot of difference in the immediate sensory inputs, according to the reliable world model. Generally speaking, however, it might prefer actions leading to high-entropy, random inputs over others—compare Sect. 3.1.

In 2005, Singh et al. (Singh et al. 2005) also used intrinsic rewards proportional to prediction errors as in Sect. 3.1 (Schmidhuber 1991c), employing a different type of reward maximizer based on the option framework which can be used to specify subgoals. As pointed out earlier, it is useful to make the conceptual distinction between the objective and the means of reaching the objective: the latter is shared

by the approaches of (Singh et al. 2005) and of Sect. 3.1; the reward maximizer is different.

In related work, Schembri et al. address the problem of learning to compose skills, assuming different skills are learned by different RL modules. They speed up skill learning by rewarding a top-level, module-selecting RL agent in proportion to the TD error of the selected module (Schembri et al. 2007).

Other researchers in the nascent field of developmental robotics (Blank and Meeden 2006; Gold and Scassellati 2006; Hart 2009; Hart et al. 2008; Kuipers et al. 2006; Olsson et al. 2006; Oudeyer and Kaplan 2006; Provost et al. 2006; Schlesinger 2006; Stronger and Stone 2006) and intrinsic reward also followed the line of basic ideas presented here, in particular, Oudeyer et al. (Oudeyer et al. 2007).

Friston et al. (2010) also propose an approach which at first glance seems similar to ours, based on free energy minimization and predictive coding. Predictive coding is a special case of compression, for example, (Schmidhuber and Heil 1996), and free energy is another approximative measure of algorithmic compressibility/algorithmic information (Li and Vitányi 1997); the latter concept is more general though. As Friston et al. write, *"Under simplifying assumptions free energy is just the amount of prediction error,"* like in the 1991 paper (Schmidhuber 1991c) discussed in Sect. 3.1. Under slightly less simplifying assumptions, it is the Kullback–Leibler divergence between probabilistic world model and probabilistic world, like in the 1995 paper (Storck et al. 1995) (which looks at the learning model before and after new observations; see Sect. 3.3). Despite these similarities, however, what Friston et al. do is to select actions that *minimize* free energy. In other words, their agents like to visit highly predictable states. Hence, their approach does *not* describe a system intrinsically motivated to learn new, previously unknown things; instead their agents really want to stabilize and make everything predictable. Friston et al. are well aware of potential objections: *"At this point, most (astute) people say: but that means I should retire to a dark room and cover my ears."* This pretty much sums up the expected criticism. In contrast, the theory of creativity has no problem whatsoever with dark rooms—the latter get boring as soon as they are predictable; then there are no wow-effects and learning progresses no more, that is, the first derivative of subjective data encodability is zero, that is, the intrinsic reward is zero, that is, the reward-maximizing agent is motivated to leave the room to find or make additional rewarding, nonrandom, learnable, novel patterns.

## *4.4 Previous, Less Formal Work in Aesthetics Theory and Psychology*

Two millennia ago, Cicero already called curiosity a "passion for learning." In the recent millennium's final century, art theorists and developmental psychologists extended this view. In its final decade, the concept eventually became sufficiently formal to permit the computer implementations discussed in Sect. 3.

### 4.4.1 Developmental Psychology

In the 1950s, psychologists revisited the idea of curiosity as the motivation for exploratory behavior (Berlyne 1950, 1960), emphasizing the importance of novelty (Berlyne 1950) and non-homeostatic drives (Harlow et al. 1950). Piaget (1955) explained explorative learning behavior in infants through his informal concepts of assimilation (new inputs are embedded in old schemas—this may be viewed as a type of compression) and accommodation (adapting an old schema to a new input—this may be viewed as a type of compression improvement). Unlike Sect. 2, however, these ideas did not provide sufficient formal details to permit the construction of artificial curious and creative agents.

### 4.4.2 Aesthetics Theory

The closely related field of aesthetics theory (Bense 1969; Birkhoff 1933; Frank 1964; Franke 1979; Moles 1968; Nake 1974) emerged even earlier in the 1930s. Why are humans somehow intrinsically motivated to observe or make certain novel patterns, such as aesthetically pleasing works of art, even when this seems irrelevant for solving typical frequently recurring problems such as hunger and even when the action of observation requires a serious effort, such as spending hours to get to the museum? Since the days of Plato and Aristotle, many have written about aesthetics and taste, trying to explain why some behaviors or objects are more interesting or aesthetically rewarding than others, for example (Collingwood 1938; Goodman 1968; Kant 1781). However, they did not have or use the mathematical tools necessary to provide formal answers to the questions above. What about more formal theories of aesthetic perception from the 1930s (Birkhoff 1933) and especially the 1960s (Bense 1969; Frank 1964; Franke 1979; Moles 1968; Nake 1974)? Some of the previous attempts at explaining aesthetic experience in the context of information theory (Bense 1969; Frank 1964; Franke 1979; Moles 1968; Nake 1974) tried to quantify the intrinsic aesthetic reward through an *"ideal"* ratio between expected and unexpected information conveyed by some aesthetic object (its *"order"* vs. its *"complexity"*). The basic idea was that aesthetic objects should neither be too simple nor too complex, as reflected by the *Wundt curve* (Wundt 1874), which assigns maximal interestingness to data whose complexity is somewhere in between the extremes. Using certain measures based on information theory (Shannon 1948), Bense (1969) argued for an ideal ratio of $1/e \sim 37\,\%$. However, these approaches were not detailed and formal enough to construct artificial, intrinsically motivated agents with a built-in desire to create aesthetically pleasing works of art.

Our formal theory of creativity does not have to postulate any objective ideal ratio of this kind. Unlike previous works emphasizing the significance of the subjective observer (Frank 1964; Frank and Franke 2002; Franke 1979), its dynamic formal definition of fun reflects the *change* in the computational resources required to encode artistic and other objects, explicitly taking into account the subjective

observer's growing knowledge as well as the limitations of its learning algorithm (or compression *improvement* algorithm). For example, random noise is always novel in the sense that it is unpredictable. But it is not rewarding since it has no pattern. It cannot be compactly encoded at all; there is no way of learning to encode it better than by storing the raw data. On the other hand, a given pattern may not be novel to a given observer at a given point in his life, because he already perfectly understands it—again there may be no way of learning to encode it even more efficiently. The value of an aesthetic experience (the fun of a creative or curious maker or observer of art) is not defined by the created or observed object per se, but by the algorithmic encoding *progress* of the subjective, learning observer.

Why did not early pioneers of aesthetic information theory put forward similar views? Perhaps because back then, the fields of algorithmic information theory and machine learning were still in their infancy?

## 5 Simple Typology of Intrinsic Motivation

By definition, intrinsic reward is something that is independent of external reward, although it may sometimes help to accelerate the latter as discussed in Sect. 3.6 (Schmidhuber 1991b, 2002a; Storck et al. 1995). So far, most if not all intrinsically motivated computational systems had the following:

1. A more or less limited adaptive encoder/predictor/compressor/model of the history of sensory inputs, internal states, reinforcement signals, and actions.
2. Some sort of real-valued intrinsic reward indicative of the learning progress of **(1)**.
3. A more or less limited reinforcement learner able to maximize future expected reward.

Hence, the typology just needs to classify previous systems with respect to properties and limitations of their specific instances of **(1–3)**. The few (if any) implementations of intrinsic motivation that do *not* fit this typology can be treated as outliers, at least until their significance/number grows, if ever. The typology actually realizes the MDL principle: find a compact model (in this case: a typology) of the data (in this case: various approaches to IM). To minimize the description length of the set of all IM approaches, minimize the description size of the typology *plus* the description size of badly modeled outliers. How? By identifying what the majority of the previous IM approaches have in common. Outliers do not fit and thus cost more extra bits, but that is ok as they are rare, since they are outliers.

**(1) Includes many subtypes characterized by the answers to the following questions:**

1. What exactly can the encoder encode (or the predictor predict or the compressor compress)?

(a) All sensory inputs as in Sect. 3.1 (Schmidhuber 1991c)? A preprocessed subset of the sensory inputs? For example, features indicating synchronicity of certain processes (Oudeyer and Kaplan 2006)? The latter may be of interest for certain limited types of IM-based learning.
(b) Reinforcement signals as in Sect. 3.1 (Schmidhuber 1991c)? (Even traditional RL agents without IM do this.)
(c) Creator actions as in Sect. 2 (Schmidhuber 2002a, 2006a, 2007c, 2009c,d)? Then even in absence of sensory feedback, curious and creative agents will be motivated to learn new motor patterns, such as previously unknown dances.
(d) Results of internal computations through sequences of internal actions as in Sect. 3.4 (Schmidhuber 2002a)? This will motivate a curious agent to create novel patterns not only in the space of sensory inputs but also in the space of abstract input transformations, such as an earlier-learned mapping from all images of cars to an internal symbol "car." The agent will also be motivated to create purely "mental" novel patterns independent of external inputs, such as number sequences obeying previously unknown mathematical laws (corresponding to mathematical discoveries).
(e) Some combination of the above? All of the above as in Sect. 3.4 (Schmidhuber 2002a)? The latter should be the default for artificial general intelligences (AGIs).

2. Is the encoder deterministic as in Sect. 3.1 (Schmidhuber 1991c), or does it predict probability distributions on possible events as in Sect. 3.3 (Storck et al. 1995)?
3. How are the encoder and its learning algorithm implemented?

(a) Is the encoder actually a continually changing, growing 3D model or simulation of the agent in the environment, used to predict future visual or tactile inputs, given agent actions (Sect. 2.2)?
(b) Is it a traditional machine learning model? A feedforward neural network mapping pairs of actions and observations to predictions of the next observation as in Sect. 3.2 (Schmidhuber 1991b)? A recurrent neural network that is in principle able to deal with event histories of arbitrary size as in Sect. 3.1 (Schmidhuber 1991c)? A Gaussian process? A support vector machine? A hidden Markov model? etc.

**(2) Includes many subtypes characterized by the answers to the following questions:**

1. Is the entire history used to evaluate the encoder's performance as in Sect. 2 (Schmidhuber 2006a, 2007c, 2009c,d) (in theory the correct thing to do, but sometimes impractical)? Or only recent data, for example, the one acquired at the present time step as in Sect. 3.2 (Schmidhuber 1991b) or in a limited time window of recent inputs? (If so, a performance decline on earlier parts of the history may go unnoticed.)
2. Which measure is used to indicate learning progress and create intrinsic reward?

(a) Mean squared prediction error or similar measures as in Sect. 3.1 (Barto et al. 2004; Klyubin et al. 2005; Schmidhuber 1991c; Singh et al. 2005)? This may fail whenever high prediction errors do not imply expected prediction progress, for example, in noisy environments, but also when the limitations of the predictor's learning algorithm prevent learning progress even in deterministic worlds.
(b) Improvements (first derivatives) of prediction error as in Sect. 3.2 (Oudeyer and Kaplan 2006; Schmidhuber 1991b)? This properly deals with both noisy/nondeterministic worlds and the computational limitations of the encoder.
(c) The information-theoretic Kullback–Leibler divergence (a.k.a. relative entropy) (Kullback 1959) between belief distributions before and after learning steps, as in Sect. 3.3 (Itti and Baldi 2005; Storck et al. 1995)? This makes sense under the assumption that all potential statistical dependencies between inputs can indeed be modeled by the given probabilistic model, which in previous implementations (Sect. 3.3) was limited to singular events (Itti and Baldi 2005; Storck et al. 1995) as opposed to arbitrary event sequences, for efficiency reasons.
(d) Minimum description length (MDL)-based measures (Rissanen 1978; Solomonoff 1964, 1978; Wallace and Boulton 1968; Wallace and Freeman 1987) comparing the number of bits required to encode the observation history before and after learning steps, as in Sect. 2 (Schmidhuber 2006a, 2007c, 2009c,d, 2010b)? Unlike the methods above, this approach automatically punishes unnecessarily complex encoders that overfit the data and can easily deal with long event sequences instead of simple one step events. For example, if the encoder uses a 3D world model or simulation, the MDL approach will ask (Sect. 2.2): how many bits are currently needed to specify all polygons in the simulation and how many bits are needed to encode deviations of the sensory history from the predictions of the 3D simulation? Adding or removing polygons may reduce the total number of bits (and decrease future prediction errors).

3. Is the computational effort of the encoder and its learning algorithm taken into account when measuring its performance, as in Sect. 2.2 (Schmidhuber 1997c, 2002a, 2006a)? The only implementation of this (Sect. 3.3; Schmidhuber 1997c, 2002a) still lacks theoretical optimality guarantees.
4. Which are the relative weights of external and intrinsic reward? This is of importance as long as the latter does not vanish in environments where after some time *nothing new* can be learned anymore.

**(3) Includes many subtypes characterized by the answers to the following questions:**

1. Which is the action repertoire of the creator?

(a) Can it produce only external motor actions, as in Sect. 3.2 (Schmidhuber 1991b; Storck et al. 1995)?

(b) Can it also manipulate an internal mental state through internal actions as in Sect. 3.4 (Schmidhuber 1997c, 2002a), thus being able to deal not only with raw sensory inputs but also with internal abstractions thereof and to create/discover novel purely mathematical patterns, like certain theoreticians who sometimes do not care much about the external world?
(c) Can it trigger learning processes by itself, by executing appropriate actions as in Sect. 3.4 (Schmidhuber 1997c, 2002a)? This is important for learning when to learn and what to learn, trading off the costs of learning versus the expected benefits in terms of intrinsic and extrinsic rewards.

2. Which are the perceptive abilities of the creator?

(a) Can it choose at any time to see any element of the entire history (Schmidhuber 2006a) of all sensory inputs, rewards, executed actions, and internal states? Or only a subset thereof, possibly a recent one, as in Sect. 3.2 (Schmidhuber 1991b)? The former should be the default for AGIs.
(b) Does it have access to the parameters and internal state of the encoder, like in Sect. 3.4 (Schmidhuber 2002a)? Or just a subset thereof? Such introspective abilities are important to predict future intrinsic rewards which depend on the already existing knowledge encoded in the encoder.

3. Which optimizer of expected intrinsic and extrinsic reward is used?

(a) A traditional Q-learner (Watkins and Dayan 1992) able to deal with delayed rewards as long as the environment is fully observable, like in Sect. 3.2? A more limited 1-step look-ahead learner (Oudeyer and Kaplan 2006) that will break down in presence of delayed intrinsic rewards? A more sophisticated RL algorithm for delayed rewards in partially observable environments (Kaelbling et al. 1996; Schmidhuber 1991d), like in Sect. 3.1? A hierarchical, subgoal-learning RL algorithm (Bakker and Schmidhuber 2004; Ring 1991, 1994; Schmidhuber and Wahnsiedler 1992; Wiering and Schmidhuber 1998a) or perhaps other hierarchical methods that do not learn to create subgoals by themselves (Barto et al. 2004; Dayan and Hinton 1993; Singh et al. 2005)?
(b) An action planner using a 3D simulation of the world to generate reward-promising trajectories (see MDL example in Sect. 2.2)?
(c) An evolutionary algorithm (Gomez et al. 2008; Holland 1975; Rechenberg 1971; Schwefel 1974) applied to recurrent neural networks (Gomez et al. 2008) or other devices that compute action sequences (Cuccu et al. 2011; Luciw et al. 2011; Schaul et al. 2011b)? Or a policy gradient method (Rückstieß et al. 2010; Sehnke et al. 2010; Sutton et al. 1999; Wierstra et al. 2008)?
(d) One of the recent universal, mathematically optimal RL algorithms (Hutter 2005; Schmidhuber 2009e), like in Sect. 2.6? Variants of universal search (Levin 1973) or its incremental extension, the Optimal Ordered Problem Solver (Schmidhuber 2004)?
(e) Something else? Obviously lots of alternative search methods can be plugged in here.

4. How does the system deal with problems of online learning?
   (a) Action sequences producing patterns that used to be novel do not get rewarded anymore once the patterns are known. Can the practical reward optimizer reliably deal with this problem of vanishing rewards, like the theoretically optimal systems of Sect. 2.6?
   (b) Can the reward optimizer actually use the continually improving predictive world model to improve or speed up the search for a better policy? This is automatically done by the above-mentioned action planner using a continually improving 3D world simulation and also by the RNN-based world model of the system in Sect. 3.1 (Schmidhuber 1991c). Does the changing model cause problems of online learning? Are those problems dealt with in a heuristic way (e.g., small learning rates) or in a theoretically sound way as in Sect. 2.6?

Each node or leaf of the typology above can be further expanded, thus becoming the root of additional straightforward refinements. But let us now address some of the recent confusion surrounding the concept of intrinsic motivation and clarify what it is *not*.

## *5.1 What IM Is Not*

### 5.1.1 Secondary Reward as an Orthogonal Issue

Reward propagation procedures of traditional RL such as Q-learning (Watkins and Dayan 1992) or RL economies and bucket brigade systems (Holland 1985; Schmidhuber 1989a,b; Wilson 1994) may be viewed as translating *rare* external rewards for achieving some goal into *frequent* internal rewards for earlier actions setting the stage. Should one call these internal "secondary" rewards intrinsic rewards? Of course not. They are just internal by-products of the method used to maximize *external* reward, which remains the only measure of overall success.

### 5.1.2 Speeding Up RL as an Orthogonal Issue

Many methods have been proposed to speed up traditional RL. Some Q-learning accelerators simply update pairs of actions and states with currently quickly changing Q-values more frequently than others (that is, Q-values with high first derivatives are favored). Others postpone updates until needed (Wiering and Schmidhuber 1998b). Again one should resist the temptation to confuse such types of secondary reward modulation with intrinsic reward, because the only thing important to such methods is the *external* reward. [Otherwise one would also have to call intrinsic reward many of the things that could be invented by any (possibly universal; Hutter 2005; Schmidhuber 2009e) RL method whose only goal is to maximize expected *external* reward.]

### 5.1.3 Subgoal Learning as an Orthogonal Issue

Some goal-seeking RL systems search a space of possible subgoal combinations, internally rewarding subsystems whose policies learn to achieve those subgoals (Bakker and Schmidhuber 2004; Ring 1994; Schmidhuber and Wahnsiedler 1992; Wiering and Schmidhuber 1998a). External reward (for reaching a final goal) is used to measure the quality of subgoal combinations: good ones survive; others are discarded. Again the internal reward for the subsystems should not be called intrinsic reward, as it is totally driven and justified by *external* reward.

### 5.1.4 Evolution of Reward Functions as an Orthogonal Issue

Essentially the same argument holds for methods that search a space of reward functions until they find one that helps a given RL method to achieve more reward more quickly, for example (Littman and Ackley 1991; Singh et al. 2009) [in many ways such methods are like the subgoal evolvers (Wiering and Schmidhuber 1998a) mentioned above]. One should not call this intrinsic reward, since once more the only thing that counts here is the *external* reward; the rest is just implementation details of the external reward maximizer.

But did not humans evolve to have this intrinsic reward component? Sure, they did, but now it is there, and now it is independent of external reward, otherwise it would not be intrinsic reward, by definition. Scientific papers on intrinsic reward should start from there. It is a different issue to analyze how and why evolution or another search process *invented* intrinsic rewards to facilitate satisfaction of *external* goals (such as survival).

## 6 How the Theory Explains Art, Science, and Humor

How does the encoding progress drive explain *humor*? Consider the following statement by comedian Bob Monkhouse:

> People laughed when I said I'd become a comedian. Well, they're not laughing now.

Some subjective observers who read this for the first time think it is funny. Why? As the eyes are sequentially scanning the text the brain receives a complex visual input stream. The latter is subjectively partially compressible as it relates to the observer's previous knowledge about letters and words. That is, given the reader's current knowledge and current encoder, the raw data can be encoded by fewer bits than required to store random data of the same size. But the punch line is unexpected for those who did not know it. Initially this failed expectation results in suboptimal data compression—storage of expected events does not cost anything, but deviations from predictions require extra bits to encode them. The encoder, however, does not stay the same forever: within a short time interval, its learning algorithm improves its performance on the data seen so far, by discovering the nonrandom, nonarbitrary,

and therefore compressible pattern relating the punch line to previous text and previous knowledge about comedians. This saves a few bits of storage. The number of saved bits (or a similar measure of learning progress) becomes the observer's intrinsic reward, possibly strong enough to motivate him to read on in search for more reward through additional yet unknown patterns.

While most previous attempts at explaining humor (e.g., Raskin 1985) also focus on the element of surprise, they lack the essential concept of *novel pattern detection* measured by compression *progress* due to learning. This progress is zero whenever the unexpected is just random white noise, and thus no fun at all. Applications of the new theory of humor can be found in recent videos (Schmidhuber 2009b).

How does the theory informally explain the motivation to create or perceive *art and music* (Schmidhuber 1997a,b, 2006a, 2007c, 2009a,c,d, 2012)? For example, why are some melodies more interesting or aesthetically rewarding than others? Not the one the listener (composer) just heard (played) twenty times in a row. It became too subjectively predictable in the process. Nor the weird one with completely unfamiliar rhythm and tonality. It seems too irregular and contain too much arbitrariness and subjective noise. The observer (creator) of the data is interested in melodies that are unfamiliar enough to contain somewhat unexpected harmonies or beats, etc., but familiar enough to allow for quickly recognizing the presence of a new learnable regularity or compressibility in the sound stream: a novel pattern! Sure, it will get boring over time, but not yet. All of this perfectly fits the principle: The current encoder of the observer or data creator tries to compress his history of acoustic and other inputs where possible. The action selector tries to find history-influencing actions such that the continually growing historic data allows for improving the encoder's performance. The interesting or aesthetically rewarding musical and other subsequences are precisely those with previously unknown yet learnable types of regularities, because they lead to encoder improvements. The boring patterns are those that are either already perfectly known or arbitrary or random or whose structure seems too hard to understand.

Similar statements hold not only for other dynamic art including film and dance (take into account the compressibility of action sequences) but also for "static" art such as painting and sculpture, created through action sequences of the artist and perceived as dynamic spatiotemporal patterns through active attention shifts of the observer. When not occupied with optimizing *external* reward, artists and observers of art are just following their encoding progress drive!

The previous computer programs discussed in Sect. 3 already incorporated (approximations of) the basic creativity principle. But do they really deserve to be viewed as rudimentary artists and scientists? The patterns they create are novel with respect to their own limited encoders and prior knowledge, but not necessarily relative to the knowledge of sophisticated adults. The main difference to human artists/scientists, however, may be only quantitative by nature, not qualitative. Current computational limitations of artificial artists do not prevent us from already using the basic principle in human–computer interaction to create art appreciable by humans—see example applications in references (Schmidhuber 1997b, 2006a, 2007c, 2009a,c,d, 2012).

How does the theory explain the nature of *inductive sciences such as physics*? If the history of the entire universe were computable and there is no evidence against this possibility (Schmidhuber 2006d), then its simplest explanation would be the shortest program that computes it. Unfortunately there is no general way of finding the shortest program computing any given data (Li and Vitányi 1997). Therefore, physicists have traditionally proceeded incrementally, analyzing just a small aspect of the world at any given time, trying to find simple laws that allow for describing their limited observations better than the best previously known law, essentially trying to find a program that encode the observed data better than the best previously known program. An unusually large encoding breakthrough deserves the name *discovery*. For example, Newton's law of gravity can be formulated as a short piece of code which allows for substantially compressing many observation sequences involving falling apples and other objects. Although its predictive power is limited—for example, it does not explain quantum fluctuations of apple atoms—it still allows for greatly reducing the number of bits required to encode the data stream, by assigning short codes to events that are predictable with high probability (Huffman 1952) under the assumption that the law holds. Einstein's general relativity theory yields additional compression progress as it compactly explains many previously unexplained deviations from Newton's predictions. Most physicists believe there is still room for further advances, and this is what is driving them to invent new experiments unveiling novel, previously unpublished patterns (Schmidhuber 2009a,c,d, 2012). When not occupied with optimizing *external* reward, physicists are also just following their encoding progress drive! All of this is compatible with Kuhn's vision of scientific revolutions (Kuhn 1962): plateaus in the evolution of a given scientific field may correspond to yet uncompressed observations that call for a new predictor or paradigm.

## 7 Concluding Remarks and Outlook

It was pointed out that systems described in the first publications on artificial curiosity and creativity (Schmidhuber 1991b,c, 2002a; Storck et al. 1995) (Sect. 3) already can be viewed as examples of implementations of a subjective data encoding progress drive that encourages the discovery or creation of novel and surprising patterns, resulting in artificial scientists or artists with various types of computational limitations, as discussed in the typology of Sect. 5. To improve previous implementations of the basic ingredients of the creativity framework and to build a continually growing, mostly unsupervised AGI, one should evaluate additional combinations of novel, advanced RL algorithms and adaptive compressors and test them on humanoid robots such as the iCub. In particular:

1. One should study better practical adaptive data encoders, such as the recent, novel artificial recurrent neural networks (RNN) (Hochreiter and Schmidhuber 1997; Schmidhuber et al. 2011) and other general yet practically feasible methods for making predictions.

2. One should reimplement the intrinsically motivated system of Sect. 3.4 (Schmidhuber 2002a) (which can learn when to learn and what to learn) with more recent, alternative, more powerful reward optimizers, such as variants of the Optimal Ordered Problem Solver (Schmidhuber 2004).
3. One should investigate under which conditions learning progress measures can be computed both accurately and efficiently, without frequent expensive compressor performance evaluations on the entire history so far.
4. Recently there has been substantial progress for a class of RL algorithms that are not quite as general as the universal ones (Hutter 2005; Schmidhuber 2002c, 2009e), but nevertheless capable of learning very general, program-like behavior in partially observable environments. One should study the applicability of recent improved RL techniques in the fields of artificial evolution (Gomez et al. 2008), policy gradients (Rückstieß et al. 2010; Sehnke et al. 2010; Sutton et al. 1999; Wierstra et al. 2008, and others).

**Acknowledgements** Thanks to Benjamin Kuipers, Herbert W. Franke, Marcus Hutter, Andy Barto, Jonathan Lansey, Michael Littman, Julian Togelius, Faustino J. Gomez, Giovanni Pezzulo, Gianluca Baldassarre, Martin Butz, Moshe Looks, Mark Ring, and several anonymous reviewers for useful comments that helped to improve this chapter or earlier papers on this subject. This research has received funds from the European Commission 7th Framework Programme (FP7/2007–2013), "Challenge 2: Cognitive Systems, Interaction, Robotics," Grant Agreement No. ICT-IP-231722, Project "IM-CLeVeR: Intrinsically Motivated Cumulative Learning Versatile Robots."

## References

Bakker, B., Schmidhuber, J.: Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In: Groen, F., Amato, N., Bonarini, A., Yoshida, E., and Kröse, B. (eds.) Proceedings of the 8th Conference on Intelligent Autonomous Systems IAS-8, pp. 438–445. IOS Press, Amsterdam (2004)

Balcan, M., .Beygelzimer, A., Langford, J.: Agnostic active learning. J. Comput. Syst. Sci. **75**(1), 78–89 (2009)

Barto, A.G., Singh, S., Chentanez, N.: Intrinsically motivated learning of hierarchical collections of skills. In: Proceedings of International Conference on Developmental Learning (ICDL). MIT, Cambridge (2004)

Bense, M.: Einführung in die informationstheoretische Ästhetik. Grundlegung und Anwendung in der Texttheorie (Introduction to Information-Theoretical Aesthetics. Foundation and Application to Text Theory). Rowohlt Taschenbuch Verlag (1969)

Berlyne, D.E.: Novelty and curiosity as determinants of exploratory behavior. Br. J. Psychol. **41**, 68–80 (1950)

Berlyne, D.E.: Conflict, Arousal, and Curiosity. McGraw Hill, New York (1960)

Birkhoff, G.D.: Aesthetic Measure. Harvard University Press, Cambridge (1933)

Blank, D., Meeden, L.: Introduction to the special issue on developmental robotics. Connect. Sci. **18**(2) (2006)

Brafman, R.I., Tennenholtz, M.: R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. J. Mach. Learn. Res. **3**, 213–231 (2002)

Cohn, D.A.: Neural network exploration using optimal experiment design. In: Cowan, J., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems 6, pp. 679–686. Morgan Kaufmann, San Francisco (1994)
Collingwood, R.G.: The Principles of Art. Oxford University Press, London (1938)
Cuccu, G., Luciw, M., Schmidhuber, J., Gomez, F.: Intrinsically motivated evolutionary search for vision-based reinforcement learning. In: Proceedings of the 2011 IEEE Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB. IEEE (2011)
Dayan, P., Hinton, G.: Feudal reinforcement learning. In: Lippman, D.S., Moody, J.E., Touretzky, D.S. (eds.) Advances in Neural Information Processing Systems 5, pp. 271–278. Morgan Kaufmann, San Francisco (1993)
Fedorov, V.V.: Theory of Optimal Experiments. Academic, New York (1972)
Frank, H.G.: Kybernetische Analysen subjektiver Sachverhalte. Verlag Schnelle, Quickborn (1964)
Frank, H.G., Franke, H.W.: Ästhetische Information. Estetika informacio. Eine Einführung in die kybernetische Ästhetik. Kopäd Verlag (2002)
Franke, H.W.: Kybernetische Ästhetik. Phänomen Kunst, 3rd edn. Ernst Reinhardt Verlag, Munich (1979)
Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: A free-energy formulation. Biol. Cybern. **102**(3), 227–260 (2010)
Gold, K., Scassellati, B.: Learning acceptable windows of contingency. Connect. Sci. **18**(2) (2006)
Gomez, F.J., Schmidhuber, J., Miikkulainen, R.: Efficient non-linear control through neuroevolution. J. Mach. Learn. Res. **9**, 937–965 (2008)
Goodman, N.: Languages of Art: An Approach to a Theory of Symbols. The Bobbs-Merrill Company, Indianapolis (1968)
Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for improved unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5) (2009)
Harlow, H.F., Harlow, M.K., Meyer, D.R.: Novelty and curiosity as determinants of exploratory behavior. J. Exp. Psychol. **41**, 68–80 (1950)
Hart, S.: The Development of Hierarchical Knowledge in Robot Systems. Ph.D. Thesis, Department of Computer Science, University of Massachusetts Amherst (2009)
Hart, S., Sen, S., Grupen, R.: Intrinsically motivated hierarchical manipulation. In: Proceedings of the 2008 IEEE Conference on Robots and Automation (ICRA), Pasadena (2008)
Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
Holland, J.H.: Properties of the bucket brigade. In: Proceedings of an International Conference on Genetic Algorithms. Lawrence Erlbaum, Hillsdale (1985)
Huffman, D.A.: A method for construction of minimum-redundancy codes. Proc. IRE **40**, 1098–1101 (1952)
Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2005). (On J. Schmidhuber's SNF grant 20-61847)
Hwang, J., Choi, J., Oh, S., II, R.J.M.: Query-based learning applied to partially trained multilayer perceptrons. IEEE Trans. Neural Netw. **2**(1), 131–136 (1991)
Itti, L., Baldi, P.F.: Bayesian surprise attracts human attention. In: Advances in Neural Information Processing Systems 19, pp. 547–554. MIT, Cambridge (2005)
Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. J. AI Res. **4**, 237–285 (1996)
Kant, I.: Critik der reinen Vernunft (1781)
Klyubin, A.S., Polani, D., Nehaniv, C.L.: Empowerment: A universal agent-centric measure of control. In: Congress on Evolutionary Ccomputation (CEC-05), IEEE (2005)
Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Prob. Inform. Transm. **1**, 1–11 (1965)
Kuhn, T.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)

Kuipers, B., Beeson, P., Modayil, J., Provost, J.: Bootstrap learning of foundational representations. Connect. Sci. **18**(2) (2006)
Kullback, S.: Statistics and Information Theory. Wiley, New York (1959)
Lenat, D.B.: Theory formation by heuristic search. Mach. Learn. **21** (1983)
Levin, L.A.: Universal sequential search problems. Prob. Inform. Transm. **9**(3), 265–266 (1973)
Li, L., Littman, M.L., Walsh, T.J.: Knows what it knows: A framework for self-aware learning. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML-08) (2008)
Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn. Springer, Berlin (1997)
Littman, M.L., Ackley, D.H.: Adaptation in constant utility non-stationary environments. In: Belew, R.K., Booker, L. (eds.) Proceedings of the Fourth International Conference on Genetic Algorithms, pp. 136–142. Morgan Kaufmann, San Mateo (1991)
Luciw, M., Graziano, V., Ring, M., Schmidhuber, J.: Artificial curiosity with planning for autonomous perceptual and cognitive development. In: Proceedings of the First Joint Conference on Development Learning and on Epigenetic Robotics ICDL-EPIROB, Frankfurt (2011)
MacKay, D.J.C.: Information-based objective functions for active data selection. Neural Comput. **4**(2), 550–604 (1992)
Moles, A.: Information Theory and Esthetic Perception. University of Illinois Press, Illinois (1968)
Nake, F.: Ästhetik als Informationsverarbeitung. Springer, Berlin (1974)
Ngo, H., Ring, M., Schmidhuber, J.: Compression progress-based curiosity drive for developmental learning. In: Proceedings of the 2011 IEEE Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB. IEEE (2011)
Olsson, L., Nehaniv, C.L., Polani, D.: From unknown sensors and actuators to actions grounded in sensorimotor perceptions. Connect. Sci. **18**(2) (2006)
Oudeyer, P.-Y., Kaplan, F.: What is intrinsic motivation? a typology of computational approaches. Front. Neurorobot. **1** (2006)
Oudeyer, P.-Y., Kaplan, F., Hafner, V.F.: Intrinsic motivation systems for autonomous mental development. IEEE Trans. Evol. Comput. **11**(2), 265–286 (2007)
Piaget, J.: The Child's Construction of Reality. Routledge and Kegan Paul, London (1955)
Plutowski, M., Cottrell, G., White, H.: Learning Mackey-Glass from 25 examples, plus or minus 2. In: Cowan, J., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems, vol. 6, pp. 1135–1142. Morgan Kaufmann, San Francisco (1994)
Poland, J., Hutter, M.: Strong asymptotic assertions for discrete MDL in regression and classification. In: Annual Machine Learning Conference of Belgium and the Netherlands (Benelearn-2005), Enschede (2005)
Popper, K.R.: All Life Is Problem Solving. Routledge, London (1999)
Provost, J., Kuipers, B.J., Miikkulainen, R.: Developing navigation behavior through self-organizing distinctive state abstraction. Connect. Sci. **18**(2) (2006)
Raskin, V.: Semantic Mechanisms of Humor. Dordrecht/Boston/Lancaster (1985)
Rechenberg, I.: Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Dissertation. Published 1973 by Fromman-Holzboog (1971)
Ring, M.B.: Incremental development of complex behaviors through automatic construction of sensory-motor hierarchies. In: Birnbaum, L., Collins, G. (eds.) Machine Learning: Proceedings of the Eighth International Workshop, pp. 343–347. Morgan Kaufmann, San Francisco (1991)
Ring, M.B.: Continual Learning in Reinforcement Environments. Ph.D. Thesis, University of Texas at Austin, Austin (1994)
Rissanen, J.: Modeling by shortest data description. Automatica **14**, 465–471 (1978)
Rückstieß, T., Sehnke, F., Schaul, T., Wierstra, D., Yi, S., Schmidhuber, J.: Exploring parameter space in reinforcement learning. Paladyn J. Behav. Robot. **1**(1), 14–24 (2010)
Schaul, T., Pape, L., Glasmachers, T., Graziano, V., Schmidhuber, J.: Coherence Progress: A Measure of Interestingness Based on Fixed Compressors. In: Fourth Conference on Artificial General Intelligence (AGI) (2011a)

Schaul, T., Sun, Y., Wierstra, D., Gomez, F., Schmidhuber, J.: Curiosity-Driven Optimization. In: IEEE Congress on Evolutionary Computation (CEC), New Orleans (2011b)

Schembri, M., Mirolli, M., Baldassarre, G.: Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In: Demiris, Y., Scassellati, B., Mareschal, D. (eds.) The 6th IEEE International Conference on Development and Learning (ICDL2007), pp. 282–287. Imperial College, London. IEEE Catalog Number: 07EX1740C, Library of Congress: 2007922394 (2007)

Schlesinger, M.: Decomposing infants' object representations: A dual-route processing account. Connect. Sci. **18**(2) (2006)

Schmidhuber, J.: A local learning algorithm for dynamic feedforward and recurrent networks. Connect. Sci. **1**(4), 403–412 (1989a)

Schmidhuber, J.: The neural bucket brigade. In: Pfeifer, R., Schreter, Z., Fogelman, Z., Steels, L. (eds.) Connectionism in Perspective, pp. 439–446. Elsevier, North-Holland, Amsterdam (1989b)

Schmidhuber, J.: Dynamische neuronale Netze und das fundamentale raumzeitliche Lernproblem. Dissertation, Institut für Informatik, Technische Universität München (1990a)

Schmidhuber, J.: An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In: Proceedings of the IEEE/INNS International Joint Conference on Neural Networks, San Diego, vol. 2, pp. 253–258 (1990b)

Schmidhuber, J.: Adaptive curiosity and adaptive confidence. Technical Report FKI-149-91, Institut für Informatik, Technische Universität München. See also Schmidhuber (1991b) (1991a)

Schmidhuber, J.: Curious model-building control systems. In: Proceedings of the International Joint Conference on Neural Networks, Singapore, vol. 2, pp. 1458–1463. IEEE (1991b)

Schmidhuber, J.: A possibility for implementing curiosity and boredom in model-building neural controllers. In: Meyer, J.A., Wilson, S.W. (eds.) Proceedings of the of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats, pp. 222–227. MIT/Bradford Books (1991c)

Schmidhuber, J.: Reinforcement learning in Markovian and non-Markovian environments. In: Lippman, D.S., Moody, J.E., Touretzky, D.S. (eds.) Advances in Neural Information Processing Systems 3 (NIPS 3), pp. 500–506. Morgan Kaufmann, San Francisco (1991d)

Schmidhuber, J.: Learning complex, extended sequences using the principle of history compression. Neural Comput. **4**(2), 234–242 (1992)

Schmidhuber, J.: Femmes fractales. Report IDSIA-99-97, IDSIA, Switzerland, (1997)

Schmidhuber, J.: Low-complexity art. Leonardo, J. Int. Soc. Arts Sci. Technol. **30**(2), 97–103 (1997b)

Schmidhuber, J.: What's interesting? Technical Report IDSIA-35-97, IDSIA (1997c). ftp://ftp.idsia.ch/pub/juergen/interest.ps.gz; extended abstract in Proc. Snowbird'98, Utah, 1998; see also Schmidhuber (2002a)

Schmidhuber, J.: Exploring the predictable. In: Ghosh, A., Tsuitsui, S. (eds.) Advances in Evolutionary Computing, pp. 579–612. Springer, Berlin (2002a)

Schmidhuber, J.: Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. Int. J. Found. Comput. Sci. **13**(4), 587–612 (2002b)

Schmidhuber, J.: The Speed Prior: a new simplicity measure yielding near-optimal computable predictions. In: Kivinen, J., Sloan, R.H. (eds.) Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002), Lecture Notes in Artificial Intelligence, pp. 216–228. Springer, Sydney (2002c)

Schmidhuber, J.: Optimal ordered problem solver. Mach. Learn. **54**, 211–254 (2004)

Schmidhuber, J.: Completely self-referential optimal reinforcement learners. In: Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S. (eds.) Artificial Neural Networks: Biological Inspirations - ICANN 2005, LNCS 3697, pp. 223–233. Springer (2005a). Plenary talk

Schmidhuber, J.: Gödel machines: Towards a technical justification of consciousness. In: Kudenko, D., Kazakov, D., Alonso, E. (eds.) Adaptive Agents and Multi-Agent Systems III (LNCS 3394), pp. 1–23. Springer, Berlin (2005b)

Schmidhuber, J.: Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. Connect. Sci. **18**(2), 173–187 (2006a)
Schmidhuber, J.: Gödel machines: Fully self-referential optimal universal self-improvers. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 199–226. Springer, Berlin (2006b). Variant available as arXiv:cs.LO/0309048
Schmidhuber, J.: The new AI: General & sound & relevant for physics. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 175–198. Springer, Berlin (2006c). Also available as TR IDSIA-04-03, arXiv:cs.AI/0302012
Schmidhuber, J.: Randomness in physics. Nature **439**(3), 392 (2006d). Correspondence
Schmidhuber, J.: 2006: Celebrating 75 years of AI - history and outlook: The next 25 years. In: Lungarella, M., Iida, F., Bongard, J., Pfeifer, R. (eds.) 50 Years of Artificial Intelligence, LNAI, vol. 4850, pp. 29–41. Springer, Berlin (2007a). Preprint available as arXiv:0708.4311
Schmidhuber, J.: New millennium AI and the convergence of history. In: Duch, W., Mandziuk, J. (eds.) Challenges to Computational Intelligence, vol. 63, pp. 15–36. Studies in Computational Intelligence, Springer, Berlin (2007b). Also available as arXiv:cs.AI/0606081
Schmidhuber, J.: Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In: Proceedings of the 10th International Conf. on Discovery Science (DS 2007), LNAI, vol. 4755, pp. 26–38. Springer, Berlin. Joint invited lecture for ALT 2007 and DS 2007, Sendai, Japan, 2007 (2007c)
Schmidhuber, J.: Driven by compression progress. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based Intelligent Information and Engineering Systems KES-2008, Lecture Notes in Computer Science, vol. 5177, Part I, p. 11. Springer, Berlin. Abstract of invited keynote (2008)
Schmidhuber, J.: Art & science as by-products of the search for novel patterns, or data compressible in unknown yet learnable ways. In: Botta, M. (ed.) Multiple ways to design research. Research cases that reshape the design discipline, Swiss Design Network - Et al. Edizioni, pp. 98–112. Springer, Berelin (2009a)
Schmidhuber, J.: Compression progress: The algorithmic principle behind curiosity and creativity (with applications of the theory of humor). 40 min video of invited talk at Singularity Summit 2009, New York City (2009b). http://www.vimeo.com/7441291. 10 min excerpts: http://www.youtube.com/watch?v=Ipomu0MLFaI.
Schmidhuber, J.: Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In: Pezzulo, G., Butz, M.V., Sigaud, O., Baldassarre, G. (eds.) Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems, LNCS, vol. 5499, pp. 48–76. Springer, Berlin (2009c)
Schmidhuber, J.: Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. J. Soc. Instrum. Control Eng. **48**(1), 21–32 (2009d)
Schmidhuber, J.: Ultimate cognition à la Gödel. Cogn. Comput. **1**(2), 177–193 (2009e)
Schmidhuber, J.: Artificial scientists & artists based on the formal theory of creativity. In: Eric, B., Marcus, H., Emanuel, K. (eds.) Artificial General Intelligence. Springer, Berlin (2010a)
Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation (1990–2010): IEEE Trans. Auton. Mental Dev. **2**(3), 230–247 (2010b)
Schmidhuber, J.: A formal theory of creativity to model the creation of art. In: McCormack, J., d'Inverno, M. (eds.) Computational Creativity. MIT, Cambridge (2012)
Schmidhuber, J., Gomez, F., Graves, A.: Sequence Learning with Artificial Recurrent Neural Networks. Cambridge University Press, Cambridge (2012) (in preparation)
Schmidhuber, J., Heil, S.: Sequential neural text compression. IEEE Trans. Neural Netw. **7**(1), 142–146 (1996)
Schmidhuber, J., Wahnsiedler, R.: Planning simple trajectories using neural subgoal generators. In: Meyer, J.A., Roitblat, H.L., Wilson, S.W. (eds.) Proc. of the 2nd International Conference on Simulation of Adaptive Behavior, pp. 196–202. MIT, Cambridge (1992)

Schmidhuber, J., Zhao, J., Wiering, M.: Shifting inductive bias with success-story algorithm, adaptive Levin search, and incremental self-improvement. Mach. Learn. **28**, 105–130 (1997)
Schwefel, H.P.: Numerische Optimierung von Computer-Modellen. Dissertation. Published 1977 by Birkhäuser, Basel (1974)
Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., Schmidhuber, J.: Parameter-exploring policy gradients. Neural Netw. **23**(4), 551–559 (2010)
Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 287–294. ACM, New York (1992)
Shannon, C.E.: A mathematical theory of communication (parts I and II): Bell Syst. Techn. J. **XXVII**, 379–423 (1948)
Singh, S., Barto, A.G., Chentanez, N.: Intrinsically motivated reinforcement learning. In: Advances in Neural Information Processing Systems 17 (NIPS). MIT, Cambridge (2005)
Singh, S., Lewis, R.L., Barto, A.G.: Where do rewards come from? In: Taatgen, N., van Rijn, H. (eds.) Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin (2009)
Solomonoff, R.J.: A formal theory of inductive inference. Part I. Inform. Control **7**, 1–22 (1964)
Solomonoff, R.J.: Complexity-based induction systems. IEEE Trans. Inform. Theory **IT-24**(5), 422–432 (1978)
Storck, J., Hochreiter, S., Schmidhuber, J.: Reinforcement driven information acquisition in non-deterministic environments. In: Proceedings of the International Conference on Artificial Neural Networks, Paris, vol. 2, pp. 159–164. EC2 & Cie (1995)
Strehl, A., Langford, J., Kakade, S.: Learning from logged implicit exploration data. Technical Report arXiv:1003.0120 (2010)
Stronger, D., Stone, P.: Towards autonomous sensor and actuator model induction on a mobile robot. Connect. Sci. **18**(2) (2006)
Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT, Cambridge (1998)
Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. ü In: Solla, S.A., Leen, T.K., Müller, K.-R. (eds.) Advances in Neural Information Processing Systems 12 [NIPS Conference, Denver, Colorado, USA, November 29 – December 4(1999)], pp. 1057–1063. MIT, Cambridge (1999)
Wallace, C.S., Boulton, D.M.: An information theoretic measure for classification. Comput. J. **11**(2), 185–194 (1968)
Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. J. R. Stat. Soc. B **49**(3), 240–265 (1987)
Watkins, C., Dayan, P.: Q-learning. Mach. Learn. **8**(3/4), 279–292 (1992)
Wiering, M., Schmidhuber, J.: HQ-learning. Adap. Behav. **6**(2), 219–246 (1998a)
Wiering, M.A., Schmidhuber, J.: Fast online Q($\lambda$): Mach. Learn. **33**(1), 105–116 (1998b)
Wierstra, D., Schaul, T., Peters, J., Schmidhuber, J.: Fitness expectation maximization. In: Proceedings of Parallel Problem Solving from Nature (PPSN 2008) (2008)
Wilson, S.: ZCS: A zeroth level classifier system. Evol. Comput. **2**, 1–18 (1994)
Wundt, W.M.: Grundzüge der Phvsiologischen Psychologie. Engelmann, Leipzig (1874)