Universal Algorithmic Ethics

Ethical Decisions based on Algorithmic Probability

Gabriel Leuenberger
March 2014

Abstract
This paper unifies the following existing formal theories: Universal Intelligence Measure, Space-time-embedded Intelligence, Observer Localisation. The unified theory provides a new simple agent framework. This new framework is then used to address philosophical questions about the simulation argument, the doomsday argument, personal identity, and ethics.

1 Prerequisites

1.1 Universal A Priori Probability
In the early days of Artificial Intelligence the pioneer Ray Solomonoff searched for a scheme to extrapolate a given bit sequence. For instance the sequence 011011011011011... should be continued with a zero. He wanted his scheme to extrapolate bit sequences of arbitrary sophistication. He solved this problem by defining the Universal A Priori Probability \( M \) also known as Solomonoff Prior. \( M(x) \) is defined as the Probability that a universal monotone Turing machine \( U \) provided with a random program will output a string that begins with \( x \). It is usually written as: \( M(x) := \sum_{p : U(p) = x} 2^{-|p|} \), where \( p \) are minimal programs for which \( U \) outputs a string \( x* \) which starts with \( x \) and \( |p| \) is the length of \( p \). The Universal Prior assigns high probabilities to “simple” strings and low probabilities to random strings. Bit sequences can be extrapolated by applying Bayesian inference to the Universal Prior, this is then called universal induction [Sol64]. For every computable probability distribution over bit strings \( x \), the expected prediction errors stay finite [Sol78][Hut01].

Universal Induction can be seen as a formalisation of Occam's Razor combined with Epicurus' principle of multiple explanations. The Philosophy of Universal Induction itself has been discussed extensively by Rathmanner and Hutter [RH11].

We will also be using the similarly defined Algorithmic Probability \( m(x) := \sum_{p : U(p) = x} 2^{-|p|} \), where \( U \) is a universal prefix Turing machine and the sum is taken over all \( p \), for which \( U \) halts and outputs \( x \). And finally we define \( m(x|y) := \sum_{p : U(p, y) = x} 2^{-|p|} \).

1.2 Universal Intelligence Measure (UIM)
By combining sequential decision theory with universal induction, Marcus Hutter invented the universally intelligent agent called AIXI [Hut05]. This lead to Shane Legg's Universal Intelligence Measure \( \Upsilon \) of a policy or agent \( a \) : \( \Upsilon(a) := \sum_{\mu \in E} 2^{-K(\mu)}, V^a_\mu \), where the sum is taken over all computable probabilistic environments \( \mu \), and \( V^a_\mu \) is the expected total utility of agent \( a \) in environment \( \mu \) [LH05][Leg08]. If \( E \) is the set of all computable probability distributions, then \( \sum_{\mu \in E} 2^{-K(\mu)} \mu(x) = \zeta(x) \approx M(x) \), where \( K \) is the Kolmogorov complexity. The main idea we take from here is the following: In order to measure the intelligence or the expected success of an agent, we assume that the probability distribution over unknowns is approximately given by the Universal Prior \( M \).
1.3 Space-time-embedded Intelligence (STEI)

The agent framework used for AIXI suffered from “unrealistic” assumptions such as perfect memory, giant amounts of computational resources and the inability to modify itself. In order to build a more realistic theory of intelligence Laurent Orseau and Mark Ring recently introduced a new agent framework called “Space-time-embedded Intelligence” [OR12]. They recursively defined the expected total utility \( V(\pi) \):

\[
V(\pi_{<t}) := \sum_{\pi_t \in H} \rho(\pi_t | \pi_{<t}) [\gamma_t u(\pi_{1:t}) + V(\pi_{1:t})],
\]

where \( \pi \) is a possible sequence of policies, \( u \) is the utility function, \( \gamma \) is the horizon function and \( \rho \) is a probability distribution function.

1.4 Observer Localization (OL)

In order to find the shortest description of our universe, multiverse theories have been developed. A fatal flaw of these theories is their lack of predictive power, due to the large number of universes. This poses an epistemological problem, which Marcus Hutter solved by showing the following:

In addition to the description of the universe (Theory of Everything) we should also include the description of how to extract our subjective experience from that computed universe. The latter is called Observer Localisation (OL). A Theory of Everything (ToE) together with an OL is called a Complete Theory of Everything (CToE). Instead of searching for the shortest ToE we should search for the shortest CToE [Hut11].

2 Unification

2.1 Interpreting STEI

In this section we are going to simplify notations of space-time-embedded intelligence slightly. In the context of a human agent, a policy \( \pi_t \) would be an encoding of the state of the human brain, detailed enough to represent the strategy of that brain.

Humans are reinforcement learning agents, where the utility \( u(\pi_{1:t}) \) equals the pleasure minus the suffering in the brain state \( \pi_t \). So it makes sense if the utility function only depends on one instead of many policies, i.e., we simplify by replacing the notation \( u(\pi_{1:t}) \) with \( u(\pi_t) \). Based on STEI: Given a sequence \( \pi \) the total utility is \( \sum \gamma_t u(\pi_t) \), where \( \pi \) is a possible sequence of policies, \( u \) is the utility function and \( \gamma \) is the horizon function. We denote the expected total utility as \( V = E[\sum \gamma_t u(\pi_t)] \).

2.2 Interpreting OL

Hutter gives the data stream coming from our visual and other sensory nerves as an example of subjective experience [Hut11]. This works well with a dualist model like AIXI. But consider a scenario where the visual cortex is damaged: Now the subjective experience is different from before, even though the visual sensory input is still the same. Another scenario: The memory is altered by an external manipulation. Now we have to say that the agent observes his own memory and has to decide, which ones of those memories are accurate and which ones are not. So the observation of his own memory is part of the subjective experience of the agent, but memory is spread throughout the entire thinking parts of the brain. Both scenarios show that the data stream coming from our sensory nerves is not a sufficient characterisation of the subjective experience of a space-time-embedded agent. Instead we should define the subjective experience as an encoding of the current state of the brain itself. This is further supported by the notion, that the brain can be decomposed into many small pieces, which can themselves be seen as agents [Min88]. So the information travelling within the brain can be seen as sensory inputs and outputs travelling between small agents, which shows the connection to Hutter's initial example.
2.3 Unifying UIM, STEI and OL
As we have seen in the previous sections, for humans, both the policy in STEI and the subjective experience in OL are encodings of the state of the brain. So for the rest of this paper we are going to treat the policy and the subjective experience as one and the same thing and call it a mind-state. To give a better description, I will borrow some words from Eray Özkalur [Özk12]: A mind-state “stores the same functional organisation of cognitive constructs”, as the brain/computer it represents. It is the “logical structure of experience”.
Analogous to how the Universal Intelligence Measure has been constructed, we now too want to use the Universal Prior $M$, in order calculate the total expected utility of an agent. As Hutter already pointed out, the CToE should be chosen according to the Occam's razor principle, which we can replace by algorithmic probability or the Universal Prior $M$. This means that the a priori probability of a CToE $p$ being correct equals $2^{-|p|}$, where $|p|$ is the length of $p$. And the a priori probability of a mind-state-sequence $x$ equals $M(x) = \sum_{p : U(p) = x^*} 2^{-|p|}$, where $p$ are minimal programs for which $U$ outputs a string $x^*$ which starts with $x$.

2.4 The new framework
The a priori probability distribution over mind-state-sequences is given by the Universal Prior. We can also express this in other words: The sequence of mind-states $\pi$ is produced by a simple universal monotone Turing machine $U$ provided with a random program, generated by fair coin flips. From now on we will use $\pi$ as a random variable for mind-state-sequences. We denote the t-th mind-state of $\pi$ as $\pi_t$.
If we additionally know one of the mind-states, we can use the Bayes theorem to update the probabilities and get an a posteriori distribution over mind-state-sequences. More formally expressed: Given the index $t$ and the mind-state $a$, using Bayes theorem, we conclude that the updated Probability of the mind-state-sequence $x$ is $P(\pi=x^* | \pi_t=a) = \frac{P(\pi=x^*) \cdot P(x_t=a)}{P(\pi_t=a)}$, where $P(\pi=x^*) = \sum_{p : U(p) = x^*} 2^{-|p|}$ and $P(\pi_t=a) = \sum_{p : |U(p)| = a} 2^{-|p|}$.

Here we assumed that $t$ was known. But the agent himself can not know for sure at which position $t$ his own current mind-state $a$ is located within this sequence, so we introduce the random variable $T$. A priori $T$ is independently distributed from $\pi$, so we need to choose the a priori distribution $P(T=t)$. We choose Algorithmic probability $m(t) =: P(T=t)$. This gives us the following probability:
$$P(\pi_T=a) = \sum_t P(T=t) \cdot P(\pi_t=a) = \sum_t m(t) \cdot \sum_{p : |U(p)| = a} 2^{-|p|} = \Theta(m(a)).$$
This is enough to address the philosophical problems from sections 3.1 to 3.3, but for the rest of the paper we need some more thoughts: Given the current mind-state $a$, the updated probability of the sequence $x$ is $P(\pi=x^* | \pi_T=a) = \sum_t P(T=t) \cdot P(\pi=x^* | \pi_t=a)$.
As in STEI we want to calculate the total expected utility $V = E_\rho \left[ \sum_t \gamma_t u(\pi_t) \right]$.

We do this using the distribution $\rho := P(\pi=x^* | \pi_T=a)$, so the total expected utility of $a$ is $V(a) := E_\rho \left[ \sum_t \gamma_t u(\pi_t) \right] = \lim_{k \to \infty} \sum_{|x|=k} P(\pi=x^* | \pi_T=a) \cdot \sum_t \gamma_t u(x_t)$.

3 Applications in Philosophy

3.1 Human mind-states
Before we address the philosophical problems of simulations, doomsdays, teleportations, and ethics,
we are going to point out some important properties of human mind-states. It has been estimated that a human brain contains around 100 Petabytes of useful information. So we assume the size of a human mind-state to be similarly large. If we randomly assemble a mind-state, then the Kolmogorov complexity of that mind-state should be expected to be very large too. But a mind-state of a human that really existed throughout history, has a much smaller Kolmogorov complexity, because its shortest description is a CToE, which is very much smaller than 100 Petabytes.

The shortest program which outputs such a human mind-state is a CToE, which can be decomposed into two parts: ToE and OL. The ToE is a program which outputs all physical events within the universe. We denote the output of the ToE as $\theta$. The OL is a program which extracts the mind-state or a mind-state sequence $\pi$ from the physical data $\theta$. This is important because algorithmic probability $m(a) = \Theta(P(\pi=a))$ is governed by the shortest programs which generate $a$. For simplicity we will only work with one $\theta$ instead of many. So for the following sections we assume $P(\pi=a) \approx g \cdot m(a) \approx M(\theta) \cdot g \cdot m(a \mid \theta)$, where $g$ is a specific constant.

### 3.2 Simulation Argument

There is the possibility that posthuman civilisations will use super computers to run realistic simulations of their ancestors, so called ancestor simulations. Nick Bostrom concluded, that if we expect, that throughout history there will be much more simulated human lives than real human lives, then we should expect, that we ourselves are living in a realistic simulation [Bos03]. Let $A$ be a mind-state random variable and let $S$ and $R$ be the sets of all mind-states of simulated humans respectively real humans, that live throughout history.

Naively one might assume $P(A \in S) \approx \frac{|S|}{|R|}$, but in this section we will derive a better formula for $P(A \in S) \approx \frac{P(A \in S)}{P(A \in R)}$ and use it, to show that Bostrom's final conclusion is not entirely correct.

For posthuman civilisations it will still be infeasible to run a perfectly accurate simulation of every physical event that occurred in their universe. So if they run a realistic ancestor simulation, it must be a computationally efficient and very sophisticated simulation. Since they do not simulate every physical event, simulated history will not be identical to real history. Therefore, given a mind-state, it would be possible to tell whether that mind-state belongs to a real human or a simulated human (assuming that there is enough computational power to do so). The simulator is part of the real universe too, so the shortest CToE that outputs mind-states of simulated humans, contains the same universe too, so the shortest CToE that outputs mind-states of simulated humans, contains the same ToE as the shortest CToE that outputs mind-states of real humans. But the OL of simulated humans can be very different from the OL of real humans.

The **substrate** is the type of physical devices, where the mind-states are located. Examples for substrates are biological brains and electronic processors. Let $Z$ be a large set of mind-states which are all located on the same substrate. Based on our new framework we get:

$$P(A \in Z) = \sum_{a \in Z} P(\pi=a) \approx \sum_{a \in Z} g \cdot m(a) \approx M(\theta) \cdot g \cdot \sum_{a \in Z} m(a \mid \theta) = M(\theta) \cdot |Z| \cdot g \cdot AM \cdot m(a \mid \theta),$$

where $AM$ is an arithmetic mean and $\theta$ is the output of the ToE.

We define the **substrate dependent factor** of $Z$ as $sdf(Z) := AM \cdot m(a \mid \theta)$. We get a better formula for the probability ratio:

$$\frac{P(A \in S)}{P(A \in R)} \approx \frac{M(\theta) \cdot g \cdot \sum_{a \in S} m(a \mid \theta)}{M(\theta) \cdot g \cdot \sum_{a \in R} m(a \mid \theta)} = \frac{|S|}{|R|} \cdot sdf(\theta) sdf(S) sdf(R)$$

In the same manner we can derive the probability that we live in a simulation:

$$P(A \in S \mid A \in S \cup R) = \frac{P(A \in S)}{P(A \in S) + P(A \in R)} \approx \frac{|S| \cdot sdf(S)}{|S| \cdot sdf(S) + |R| \cdot sdf(R)}$$

It could be that the sdf of biological brains is much larger than the sdf of computationally efficiently simulated brains. In this case the large numbers of simulated people could be compensated by the small sdf of simulated people. Depending on the computer architecture, also the opposite could hold true. I leave it to future papers to determine the ratios between sdf of different substrates.
Philosopher Bostrom concluded, that at least one of the following three propositions is true:

1. *Our civilization and others will almost certainly go extinct before reaching posthuman stage.*
2. *Posthuman civilisations run almost no ancestor simulations, because they are not interested.*
3. *Almost all civilisations are simulated and we almost certainly live in a simulation too.* [Bos03]

But now, based on our formulas we see, that we have to add more possibilities:

4. *Almost all civilisations are simulated, but we probably do not live in a simulation, because of the extremely small ratio between the sdf of very efficiently simulated brains and the sdf of full-sized biological brains.*
5. *Almost all civilisations are real, but we probably are simulated, because of the extremely large ratio between the sdf.*

When certain features of the ToE are known one could try to calculate an sdf by estimating the minimal length of required OL programs. The sdf might be correlated with the average spatial volume of a mind of that substrate.

### 3.3 Doomsday Argument

When working on the doomsday argument, Bostrom stated the Strong Self-Sampling Assumption: “Each observer-moment should reason as if it were randomly selected from the class of all observer-moments in this reference class.” [Bos02] We change this to: Each observer-moment (or mind-state) \( a \) should reason as if it were randomly selected with the probability \( P(\pi_T = a) \).

Based on this, it should be decided whether the doomsday argument is valid or not. I suspect it to be invalid, because the probability that \( a \) is human, might scale approximately linearly with the number of real human mind-states.

### 3.4 Personal Identity

It is an equivalence relation which tells us whether two entities are the same person or not. Derek Parfit used a teleportation thought experiment to discuss the properties of personal identity. A teleporter is a device, which scans your body, while disassembling it fast enough, so you do not notice anything. The scan information is then sent to the destination and used to reassemble your body out of new matter. In his thought-experiment, using this technology, Parfit travels to Mars at the speed of light. The following text is a quote from his philosophy book [Par84]:

“Simple Teletransportation, as just described, is a common feature in science fiction. And it is believed by some readers of this fiction, merely to be the fastest way of travelling. They believe that my Replica would be me. Other science fiction readers and some of the characters in this fiction take a different view. They believe that, when I press the green button, I die. My Replica is someone else, who has been made to be exactly like me.”

So the main question here is, whether teleportation kills people or not.

In this section we will formalize this question based on our new framework. Recall the definition of the total expected utility \( V(a) := E_\rho \left[ \sum \gamma_i u(\pi_i) \right] \).

Given the current mind-state \( a \) we now define the weight of a mind-state \( b \) as:

\[
  w(b|a) := E_\rho \left[ \sum \gamma_i [\pi_i=b] \right] = \lim_{k \to \infty} \sum_{|x|=k} P(\pi=x* | \pi_T = a) \cdot \sum \gamma_i [x_i = b]
\]

Now we can write the total expected utility of \( a \) more elegantly:

\[
  V(a) = \sum_b w(b|a) \cdot u(b)
\]

Parfit argued that personal identity should be a gradual function instead of an equivalence relation [Par84]. Because \( w(b|a) \) tells us how much we should value the utility of a certain mind-state \( b \), the function \( w(b|a) \) already is the right function to represent personal identity.

In order to calculate if teleportation kills, we have to compare the following two scenarios:

- **Tele**: \( A \) is the mind-state of a human one day before being teleported and \( B \) is the mind-state of that same human one day after being teleported.
- **Stay**: \( A \) and \( B \) are mind-states of the same human two days apart and he was not teleported within those two days.
Following ratio tells us how much personal identity is lost through teleportation: $\frac{E_{\text{Tel} \{ w(B \mid A) \}}}{E_{\text{Stay} \{ w(B \mid A) \}}}$

-If this ratio is close to one, the teleporter actually transports persons to Mars.
-If this ratio is close to zero, the teleporter just kills persons and creates new persons.

It is unlikely that there ever will be teleporters for humans, but this theory will become relevant, as soon as there will be intelligent agents on electronic devices, which can travel to an other place, using the internet as their teleporter. The theory will also be relevant, when they will be making multiple copies of themselves.

## 4 Universal Algorithmic Ethics

### 4.1 Weight and Similarity

A horizon function $\gamma_i$ still needs to be chosen. For maximal farsightedness one could choose universal discounting $\gamma_i = 2^{-K(i)}$ as suggested by Hutter [Hut05 page 170]. We will use a slightly different version of universal discounting: $\gamma_i := m(t)$, it inherits the main properties of the original version. Özkural suggested to use mutual algorithmic information $I_k(a \mid b)$ to describe the similarity between two cognitive systems [Özk12]. A similar approach would be, to use the Universal similarity metric aka normalized information distance $\text{NID}(a,b)$ [LCL04]. In this section we show that the weight $w(b \mid a)$ is roughly proportional to $m(b \mid a)$, and show how this relates to mutual algorithmic information and to the universal similarity metric.

$$w(b \mid a) := E_{\rho} \{ \sum_i \gamma_i \{ \pi_i = b \} \} = \lim_{k \to \infty} \sum_{|x|=k} P(\pi=x \mid \pi_i = a) \cdot \sum_i \gamma_i \{ x_i = b \}$$

$$= \lim_{k \to \infty} \sum_{|x|=k} (\sum_i P(T=i) \cdot P(\pi=x \mid \pi_i = a) \cdot \sum_i \gamma_i \{ x_i = b \})$$

$$= \lim_{k \to \infty} \sum_{|x|=k} (\sum_i P(T=i) \cdot P(\pi=x \mid \pi_i = a) \cdot \gamma_j \{ x_j = b \})$$

$$= P(\pi=a)^{-1} \cdot \lim_{k \to \infty} \sum_{|x|=k} P(\pi=x) \cdot \sum_{(i,j)} P(T=i) \cdot \gamma_j \{ (x_i, x_j) = (a,b) \}$$

$$= P(\pi=a)^{-1} \cdot \lim_{k \to \infty} \sum_{|x|=k} M(x) \cdot \sum_{(i,j)} m(i) \cdot m(j) \cdot \gamma_j \{ (x_i, x_j) = (a,b) \}$$

$$\approx P(\pi=a)^{-1} \cdot \lim_{k \to \infty} \sum_{|x|=k} M(x) \cdot m(a \mid x) \cdot m(b \mid x) \cdot h$$

$$\approx P(\pi=a)^{-1} \cdot \gamma \cdot m(a,b) \approx \frac{g \cdot h \cdot m(a,b)}{g \cdot m(a)} = \frac{h \cdot m(a,b)}{m(a)} \propto m(b \mid a) = \Theta(2^{-K(b \mid a)})$$

where $h$ is a specific constant. We can use this result to show the relation of the weight $w$ to the similarity measures:

$$- \log \{ w(b \mid a) \cdot w(a \mid b) \} = \Theta( K(b \mid a) + K(a \mid b) ) = \Theta( K(a,b) - I_k(a \mid b) )$$

$$- \log \{ w(b \mid a) \cdot w(a \mid b) \} = \Theta( K(b \mid a) + K(a \mid b) ) = \Theta( K(a,b) \cdot \text{NID}(a,b) )$$

where $I_k(a \mid b)$ is the mutual algorithmic information and $\text{NID}(a,b)$ is the normalized information distance. We conclude, that the amount of similarity between two mind-states is positively correlated with the weights, that they assign to each other.

### 4.2 Resolution Independence

We used a ratio between weights to address the philosophical problem in section 3.4. and we will be using such ratios in the next section as well. So here we are going to discuss an important property of ratios between weights of the form $w(c \mid a) / w(b \mid a)$. When we introduced the concept of mind-states, we did not exactly specify how physically accurate or how abstract such representations
We summarize our abstraction of Universal Algorithmic Ethics in the following two lines:

utilities of his own future mind-state

If a

Since

those mind-states are produced by a simple universal Turing machine with a random program.

This is a theory of ethics which we get as a by-product of our simple initial assumption that mind-states are mind-states at a very high resolution, e.g. molecular level, and let \( a', b' \) and \( c' \) be the same mind-states, but at a much lower resolution, e.g. mentalese level.

Those mind-states correspond such that \( m(a' | a) = m(b' | b) = m(c' | c) \) holds, because the low-res mind-states can be computed based on the corresponding high-res mind-states, using the same program for all three cases. Based on this equation and given that the low-res mind-states are distinct enough amongst each other, we get the following equation, where \( f \) is some constant:

\[
m(a', b' | a, b) = m(a' | a) \cdot m(b' | b) \cdot f = m(a' | a) \cdot m(c' | c) \cdot f = m(a', c' | a, c)
\]

Using this equation and the approximation \( w(b | a) \approx \frac{h \cdot m(a, b)}{m(a)} \) derived in section 4.1, we get the following relation between the weight ratios of different resolutions:

\[
\frac{w(c | a)}{w(b | a)} \approx \frac{h \cdot m(a, b)}{m(a)} \cdot \frac{m(a)}{m(a, c)} = \frac{m(a, b | a, c)}{m(a, c)} = \frac{m(a, c' | a, c)}{m(a', c' | a, c)} \approx \frac{w(c' | a')}{w(b' | a')}
\]

The just derived relation \( \frac{w(c | a)}{w(b | a)} \approx \frac{w(c' | a')}{w(b' | a')} \) suggests that there is a large range of different resolutions, which all result in roughly the same weight ratios. As we will see in the next section, decisions of rational agents strongly depend on such weight ratios, and are therefore resolution-independent for a large range of resolutions.

### 4.3 Rational Game Strategies

We call an agent rational if he aims towards maximizing the total expected utility

\[
V(a) = \sum_b w(b | a)u(b)
\]

The weight \( w(b | a) \) tells us how much we should value the utility of another mind-state \( b \), while having the mind-state \( a \). We could also use the weight to calculate how much we should value the utilities of mind-states of agents other than ourselves.

This is a theory of ethics which we get as a by-product of our simple initial assumption that mind-states are produced by a simple universal Turing machine with a random program. Analogous to the names Universal Induction, Universal Algorithmic Intelligence, Universal Intelligence Measure, and Universal Similarity Measure, we call our theory Universal Algorithmic Ethics. It is universal in the same sense that those other theories are universal and holds across different resolutions. In this section we introduce a useful abstraction of Universal Algorithmic Ethics and provide an example of how game theory can be transformed. We introduce the human mind-state random variables \( A, B, C \in R \). They are not independently distributed as described in the following two sentences.

\( A \) and \( B \) are mind-states of the same human at different points in time.

\( B \) and \( C \) are mind-states of different humans at the same point in time.

We define \( A := E \left[ \frac{w(C | A)}{w(B | A)} \right] \). I leave open how exactly \( A, B, C \) should be distributed, since the choice of the distribution probably does not influence \( A \) very much. The simplest would be a uniform distribution over all possible triples which are consistent with the sentences above.

\( A \) can be seen as a constant value of humankind which should be determined in future papers.

Since \( A \) is based on a fraction between weights, it would also be resolution independent.

If a rational human is taking a decision while his mind-state is \( A \) and this decision will affect the utilities of his own future mind-state \( B \) and someone else's future mind-state \( C \), then he would take an egoistic decision if \( A \) is close to zero, or take a utilitarian decision if \( A \) is close to one.

We summarize our abstraction of Universal Algorithmic Ethics in the following two lines:
If $\Lambda$ is very close to zero, then human egoism is more rational.

If $\Lambda$ is close to one, then human utilitarianism is more rational.

Now imagine a game with two players. At the end of the game the players get the pair of rewards $(r_B, r_C)$, which depend on the decisions the players took. The first agent gets $r_B$ and the second agent gets $r_C$. In classical game theory both agents act purely egoistically in order to maximize their own reward. But what if they act according to Universal Algorithmic Ethics? Here we show how to transform a game by transforming the reward pairs:

$$(r'_B, r'_C) = (r_B, r_C) \cdot \left( \frac{1-\vartheta}{\vartheta}, \frac{\vartheta}{1-\vartheta} \right),$$

where

$$\vartheta := E \left[ \frac{w(C | A)}{w(B | A) + w(C | A)} \right] = \frac{1}{1+\Lambda}.$$

This transformation of the reward pairs makes rational agents behave according to Universal Algorithmic Ethics. The transformation maintains the sum between the two rewards, i.e., the total amount of available rewards in the game stays the same. This is useful in order to compare the performance in the transformed game to the performance in the original game. We illustrate this by taking the following prisoners dilemma as example (numbers are years in prison):

|                | cooperation | defection |
|----------------|-------------|-----------|
| cooperation    | (1, 1)      | (5, 0)    |
| defection      | (0, 5)      | (4, 4)    |

**Example : Original Prisoners dilemma**

**How to transform the Prisoners dilemma**

| $\vartheta$ | cooperation | defection |
|-------------|-------------|-----------|
| 0.0         | (1, 1)      | (5, 0)    |
|             | (0, 5)      | (4, 4)    |

| $\vartheta=0.0$ | cooperation | defection |
|-----------------|-------------|-----------|
| cooperation     | (1, 1)      | (5, 0)    |
| defection       | (0, 5)      | (4, 4)    |

**Dominating Strategy:** Defection

**Transformed Prisoners dilemma $\vartheta=0.0$**

| $\vartheta=0.2$ | cooperation | defection |
|-----------------|-------------|-----------|
| cooperation     | (1, 1)      | (4, 1)    |
| defection       | (1, 4)      | (4, 4)    |

**Dominating Strategy:** Both

| $\vartheta=0.4$ | cooperation | defection |
|-----------------|-------------|-----------|
| cooperation     | (1, 1)      | (3, 2)    |
| defection       | (2, 3)      | (4, 4)    |

**Conclusion:** If $\vartheta$ is large enough, the rational prisoners cooperate. In this case the prisoners dilemma is not a dilemma any more. The summed prison time is minimized.

### 4.4 Outlook and Conclusion

How the utility function works has been left open by this paper. A few suggestions can be made:

The mind-states might be written in a language specialized to represent mind-states, so the utility function could directly read out and sum up the utility found within a mind-state. Another possibility would be to use a universal utility function, which does not need to know the language, but instead uses a form of universal induction to infer which utility function an agent is trying to maximize. It would be biased towards utility functions which can be calculated by short programs. Since humans are reinforcement learning agents, it is most straightforward to define the utility as pleasure minus suffering, but other utility functions should be considered as well. For simplicity within this paper, the whole human brain was treated as one decision making entity with full internal cooperation between the different brain parts. But if different brain parts of a brain don't perfectly cooperate with each other, because of their conflicting objectives, they should be viewed as separate entities. Each brain part having its own mind-state and taking its own decisions.
The value of $\Lambda$ could be close to zero or close to one. According to the principle of indifference I suggest that $\Lambda = 0.5$ should be used until evidence for either of the two cases is found.

In this paper we developed a universal and resolution independent theory of ethics (or morality) based on the principles of algorithmic probability theory. In a future where digitalized minds can be copied, sped up, modified, sent around, and maybe even merged, everyone will be forced to think about new theories of epistemology, personal identity, and ethics. With this paper the first step in this direction has been taken.

Acknowledgements
Thanks to Eray Özkural for conversations on philosophy and algorithmic probability.
Thanks to Lukas Gloor for sharing his views on personal identity and ethics.

References:
[Sol64] R. J. Solomonoff. *A Formal Theory of Inductive Inference.* Information and Control 7, p. 1-22, 1964.
[Sol78] R. J. Solomonoff. *Complexity based induction systems: comparisons and convergence theorems.* IEEE Transactions on Information Theory, p. 422-432, 1978.
[Hut01] M. Hutter. *Convergence and Error Bounds for Prediction of Nonbinary Sequences.* Proceedings twelfth European Conference on Machine Learning, p. 239-250, 2001.
[RH11] S. Rathmanner and M. Hutter. *A Philosophical Treatise of Universal Induction.* Entropy 13, p. 1076-1136, 2011.
[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Information.* Springer, Berlin, 2005.
[LH05] S. Legg and M. Hutter. *A Universal Measure of Intelligence for artificial agents.* International Joint Conference On Artificial Intelligence Vol. 19, p. 1509, 2005.
[Leg08] S. Legg. *Machine Super Intelligence.* Department of Informatics, University of Lugano, 2008.
[OR12] L. Orseau and M. Ring. *Space-time-embedded Intelligence.* Artificial General Intelligence 5, p. 209-218, 2012.
[Hut12] M. Hutter. *The Subjective Computable Universe.* A computable Universe: Understanding & Exploring Nature as a Computation, p. 399-216, 2012.
[Min88] M. Minsky. *Society of Mind.* Simon and Schuster, New York, 1988.
[Özk12] E. Özkural. *What is it like to be a Brain Simulation?*. Artificial General Intelligence 5, p. 232-241, 2012
[Bos03] N. Bostrom. *Are we living in a computer simulation?*. The Philosophical Quarterly Vol. 53, p. 243-255, 2003.
[Bos02] N. Bostrom. *Anthropic Bias: Observation Selection Effects in Science & Philosophy.* Routledge, New York, 2002.
[Par84] D. Parfit. *Reasons and Persons* chapter ten and chapter fourteen. Clarendon Press, Oxford, 1984.
[LCL04] M. Li, X. Chen, X. Li, B. Ma, P. M. B. Vitany. *The Similarity Metric.* IEEE Transactions on Information Theory, Volume XX, p. 3250-3264, 2004.