Formalizing Preference Utilitarianism in Physical World Models

Caspar Oesterheld
caspar.oesterheld@uni-bremen.de

April 23, 2015

Abstract

Most ethical work is done at a low level of formality. This makes practical moral questions inaccessible to formal and natural sciences and can lead to misunderstandings in ethical discussion. In this paper, we use Bayesian inference to introduce a formalization of preference utilitarianism in physical world models, specifically cellular automata. Even though our formalization is not immediately applicable, it could provide ethics and ultimately the question of how to “make the world better” with a formal basis.

Keywords: preference utilitarianism, formalization, artificial life, (machine) ethics

1 Introduction

Usually, ethical imperatives are not formulated with sufficient precision to study them and their realization mathematically. (McLaren, 2011, p. 297; Gips, 2011, p. 251) In particular, it is impossible to implement them on an intelligent machine to make it behave benevolently in our universe, which is the subject of a field known as Friendly AI (e.g., see Yudkowsky, 2001, p. 2) or machine ethics (e.g., see M. Anderson and S. L. Anderson, 2011, p. 1). Whereas existing formalizations of utilitarian ethics have been successfully applied to economics, they are incomplete due to the nature of their dualistic world model in which agents interact with their environment only through a designated interface.

In this paper however, we take the following steps towards a workable and simple formalization of preference utilitarianism in physical world models:

- We describe the problem of informality in ethics and the shortcomings of previous dualist approaches to formalizing utilitarian ethics (section 2)

\footnote{For introductions to and ethical discussions of the underlying notion of preference utilitarianism see Tomasik (2015a) and Tomasik (2015b).}
• We justify cellular automata as a world model, use Bayes' theorem to extract utility functions from a given space-time embedded agent and introduce a formalization of preference utilitarianism (section 3).

• We compare our approach with existing work in ethics, game theory and artificial intelligence (section 4). Our formalization is novel but nevertheless relates to a growing movement to treat agents as embedded into the environment.

2 The problem of formalizing ethics in physical systems

Discussion on informally specified moral imperatives can be difficult due to different interpretations of the texts describing the imperative. Thus, formalizing moral imperatives could augment informal ethical discussion. (Gips, 2011, p. 251; S. L. Anderson, 2011; Dennett, 2006; Moor, 2011, p. 19)

Furthermore, science and engineering answer formally described questions and solve well-specified tasks, but are not immediately applicable to the informal question of how to make the world “better”.

This problem has been identified in economics and game theory, which has led to some very useful formalizations of utilitarianism (e.g. Harsanyi, 1982).

However, their formalization relies on consciousness-matter dualism: They are based on the idea of non-physical agents interacting with the physical environment through a designated interface. The agents are not part of the physical world or embedded into it, so that their thoughts or computations can not be influenced by physical laws. Also, agents’ utility functions are assumed to not depend on the agents (or their physical configurations) themselves. These are typical assumptions in game theory. After all, game theory is about games, in which players are not actually inside the game, nor can they decide themselves what goals to pursue. This classic (multi-)agent-environment model is depicted in figure [1].

Our world, however, is (usually presumed to be) a purely physical system: Ethically relevant entities (animals etc.) are embedded in the environment. For example, our brains behave according to the same laws of physics as the rest of the world and there is no designated interface between our mind and the physical world. Also, happiness and preferences are not given by predetermined utility functions or rewards from the environment, but are the result of physical processes in our bodies. Therefore, dualist descriptions and formalizations leave questions unanswered:

• What objects are ethically relevant? (What are the agents of our non-dualist world?)

• What is a space-time embedded agent’s or, more generally, an object’s utility function?
Thus, even though classic formalizations of (preferentist) utilitarianism in the agent-environment-model can formalize the vague notions of goals and preferences with utility functions, these formalizations are incomplete, at least in our physical, non-dualist world.

3 A Bayesian approach to formalizing preference utilitarianism in physical systems

3.1 Cellular automata as non-dualist world models

To overcome the described problems of dualist approaches to utilitarianism, we first have to choose a new, physical setting for our ethical imperative. Instead of employing string theory and other contemporary theoretical frameworks, we choose a model that is much more simple to handle formally: cellular automata. These have sometimes even been pointed out to be candidates for modeling our own universe, (ch. 9 Wolfram, 2002; Schmidhuber, 1999; Zuse, 1967; Zuse, 1969) but even if physics will prove cellular automata to be a wrong model, they may still be of instrumental value for the purpose of this paper. (compare Downey, 2012, pp. 70f., 77-79; Hawking and Mlodinow, 2010, ch. 8)

For detailed introductions to classic cellular automata with neighbor-based rules, see Wolfram (2002) or Shiffman (2012, ch. 7) for a purely informal and Wolfram (1983) for a slightly more technical treatment that focuses on one-dimensional cellular automata. In section 3.1.1 we will consider a generalized and relatively simple formalism, which is not limited to rules that only depend on neighbors of a cell.
In CA, it is immediately clear that for a (preference) utilitarian morality we have to answer the questions that are avoided by assuming a set of agents and their utility functions to be known from the beginning. It also frees us from many ethical intuitions that we build up specifically for our own living situations and reduces moral intuition to its very fundamentals.

Figure 2 shows a state of a cellular automaton illustrating the problem of defining utilitarianism or any other ethical imperative in physical models. Clearly, many intuitions are very difficult (if not impossible) to formulate universally and precisely in CA. Thereby, the required formality helps in choosing and defining an ethical imperative.

3.1.1 A formal introduction to cellular systems

We now introduce some very basic notation and terminology of cellular systems, a generalization of classic cellular automata, thus setting the scene for our ethical imperative.

For given sets $A$ and $B$, let $A^B$ denote the set of functions from $B$ to $A$. A cellular system is a triple $(C, S, d)$ of a countable set of cells $C$, a finite set of cell states $S$ and a function $d : S^C \rightarrow S^C$ that maps a world state $s : C \rightarrow S$ onto its successor. So basically a world consists of a set of cells that can have different values and a function that models deterministic state-transitions.\footnote{The choice of deterministic systems was made primarily to simplify the formalization. It appears to be unproblematic to transfer formal preference utilitarianism to non-deterministic systems, but defining non-deterministic cellular automata themselves is a little more difficult.}

Cells of cellular systems do not necessarily have to be on a regular grid and computing new states does not have to be done via neighbor-based lookup tables.
This makes formalization much easier.

But before anything else, we have to define structures which represent objects in our cellular systems. A space $Spc \subseteq C$ in a cellular system $(C, S, d)$ is a finite subset of the set of cells $C$. A structure $str$ on a space $Spc$ is a function $str: Spc \rightarrow S$ that maps the cells of the space onto cell values.

A history is a function $h: \mathbb{N} \rightarrow S^C$ that maps natural numbers as time steps onto states of the system. For example, the history $h_s$ of an initial state $s$ can then be defined recursively by $h_s(n) = d(h_s(n-1))$ for $n \geq 1$ with the base case $h_s(0) = s$.

3.2 Posterior probabilities and the priority of a (given) goal to a given agent

Now, the setting is right, so that we can ask the question: Does a particular object or structure in a cellular automaton want anything? I.e. does it have relevant preferences? Using the formal model of utility functions for preferences, does a particular structure want to maximize some function $u: (S^C)\mathbb{N} \rightarrow \mathbb{R}$ that maps histories of the world onto utilities?

It is fruitful to think about how one would approach such questions in our world, when encountering some very odd organism. At least one possible approach would be to put it into different situations or environments and see what it does to them. If the structure increases some potential utility function in different environments, it seems as if this utility function represents an aspect of the structure’s preferences.

However, for some utility functions it is not very special that their values are increased and then it might just be coincidence that the structure in question also does so. For example, it is usually not considered a structure’s preference to increase entropy even if entropy increases in environments including this structure, because an increase in entropy is extremely common with or without the structure.

Also, we feel that some utility functions are less likely than others by themselves, e.g. because they are very complex or specific.

But how can we formally capture these notions?

Since uncertainty is involved, we interpret the degree to which a utility function $u$ is important to some structure $str$ that exists at time step $i$ as the posterior probability of that utility function given the structure, a probability we denote by $P(u|str@i)$, where $str@i$ denotes the event that $str$ exists in time step $i$. Here, the utility function is interpreted as a hypothesis about the structure’s

\footnote{Other codomains of utility functions seem possible as long as they are subsets of a totally ordered vector space over $\mathbb{R}$. Examples are $[-1,1]$ or real numbers with some infinities $\mathbb{R} \cup \{-\infty\}$.}

\footnote{Alternatively, one can try to avoid this hypothetical experiment by predicting the organism’s behavior. For example, one could try to ask the organism what it would do or infer its typical behavior from its internals.}

\footnote{Including $i$ into the data is important, because otherwise identical structures at different points in time would have identical utility functions. This is a problem, when the utility function $u$ is applied to the whole history, because then structures cannot have preferences...}
true intentions. In a purely physical, non-dualist world there is nothing but the structure itself, of course. Therefore, the “true intentions” do not really exist, which makes it still hard to know what \( P(u|str@i) \) is supposed to mean. To finally overcome this problem, we will equate intention and purpose, i.e. we equate the following interpretations of \( u \) as a hypothesis explaining the data \( str@i \): (compare Dennett, 1989, pp. 289ff., 299f., 318, 320f.)

- The structure \( str \) has the utility function \( u \) as one of its goals.
- Maximizing \( u \) was a goal of an entity that chose \( str \).

The second interpretation is more useful, because it describes a data-generating process and thus comes closer to typical statistical models. Thus, we have to find the posterior probability of some model (a utility function) given some data (a structure). For this problem Bayes’ theorem suggests itself, because it provides an equation for posterior probabilities. In our case, Bayes’ theorem can be used to infer the likelihood that some utility function was a goal when a structure was chosen from some priors and the likelihood of choosing the structure given that the goal is to maximize the utility function. Specifically, Bayes’ theorem gives us

\[
P(u|str@i) = \frac{P(str@i|u) \cdot P(u)}{P(str@i)},
\]

where \( P(u) \) and \( P(str@i) = P(str) \) are prior probability distributions of utility functions and structures, respectively, and \( P(str@i|u) \) is the probability of (some hypothetical entity choosing) \( str \) at time step \( i \) when \( u \) is to be maximized. Whereas \( P(u|str@i) \) is very hard to grasp intuitively, it is more clear what the probability distributions on the right hand side of the equation mean. Nevertheless, they do not correspond to measurable probability distributions like the results from rolling a dice. Indeed, \( P(u) \) and \( P(str) \) are ultimately subjective (e.g. see Olshausen, 2004, pp. 1f. Robert, 1994, p. 9) and \( P(str@i|u) \) depends on what exactly the hypothesis \( u \) is supposed to express, thus leaving our ethical imperative parametrized by these distributions.

Nonetheless, there seem to be canonical approaches. The ratio \( P(str@i|u)/P(str) \) can be interpreted as the probability of choosing \( str \) when maximizing \( u \) without considering differences in the a priori probabilities of different structures. And it may be defined in proportion to the performance of \( str \) measured by \( u \) averaged over different environments. The probability

about themselves (“personal happiness”) without also having preferences about all other identical structures (at the same place). An alternative would be to apply utility functions only to the part of the history from the point of the existence of the structure onwards, so that identical structures at different points in time have equal utility functions that are applied differently. However, it seems like this neglects that the past can depend on the action of an agent in the present, as illustrated in Newcomb’s paradox by Nozick (1969).

Note that some structures, like humans, shall have many goals at the same time and some, like unstable pieces of matter, (almost) none. Therefore, utility functions are not meant to be mutually exclusive or even collectively exhaustive. So, we do not require the sum of probabilities over all utility functions \( \sum_u P(u|str@i) \) to be 1. Instead, every non-negative real number is possible.
distribution of utility functions on the other hand seems to be well suited for Solomonoff’s prior based on Kolmogorov complexity (e.g. see Legg, 1997).

Finally, note how Bayes’ theorem catches our intuitions from above, especially when assuming probability distributions similar to the suggested ones: When some structure \( str \) maximizes some utility functions \( u \) very well, then \( P(str @ i | u) \) and thereby the relevance of the utility function to the object would increase. On the other hand, if many other structures are comparably good, then the probability for each one to be chosen when given the utility function is smaller (due to the sum of the probabilities of all possible structures being 1) and the probability of the utility function being a real preference would decrease with it. Finally, multiplying by \( P(u) \) catches abstruse utility functions, e.g. utility functions that are specifically suited to be fulfilled by the structure in question.

3.3 An individual structure’s welfare function

Having introduced a way of determining how likely it is that some utility function is the utility function of some object, we define the welfare \( U_{str @ i} \) of a structure \( str \) that exists at some step \( i \) of a history \( h \), as the weighted sum over all utility functions

\[
U_{str @ i} = \sum_u P(u | str @ i) u(h),
\]

(2)

where \( h \) is the history and the sum is over all theoretically possible utility functions \( u : (S^C)^N \rightarrow \mathbb{R} \).

We call this term expected utility, because this expression is generally used for adding utilities based on their likelihood, which is a common concept. However, the term usually suggests that there is also an actual utility. In our case of ascribing preferences to physical objects however, no such thing exists. We only imagine there to be some real utility or welfare functions and that we use Bayesian inference to find them. But in fact, the structure itself is all there exists and thus the expected utility is as actual as possible.

The sum in the term for expected utility is over an uncountably infinite set, which can only converge when only countably many summands are non-zero.\(^7\) Some other concerns are described in footnote \(^9\) and addressed in footnote \(^{10}\).

3.4 Summing over all agents

The utilitarian imperative is to maximize a global welfare function that is the sum of all individuals' welfare functions. We already defined the welfare function of single structures. So next we have to define what the set of all agents is and how to sum over it. As foreshadowed before, we will consider all possible structures of a cellular automaton using equation \(^2\) and rely on (intuitively) irrelevant ones to receive very low \( P(u | str @ i) \) values. To sum the utility over all agents, we not only have to sum over all structures in a particular state, but

\(^7\) If Solomonoff’s prior is chosen for \( P(u) \), all incomputable utility function have zero probability. Since the set of computable functions is countable, only countably many summands could possibly be non-zero.
first over all (discrete) time steps of the history of the cellular automaton world and only then over all structures in every state. This way, we sum the welfare of all agents ever coming into existence. For the summands, we can insert the term obtained in equation 2

\[ \sum_i \sum_{\text{str} \in \text{Sp}} \sum_u P(u|\text{str} \in i) u(h), \] (3)

where \( U_{\text{str} \in i} \) denotes the welfare or utility of the structure \( \text{str} \) that exists at time step \( i \), the first sum is over all integers functioning as time steps, the second is over all structures in \( h(i) \) and the third over all possible utility functions.\footnote{More precisely, but less elegantly, one could write}

\[ \sum_{i=0}^{\infty} \sum_{\text{Sp} \in \text{Fin}(C)} \sum_u u(h) P(u|\text{Sp} \in i) \] \( \) where \( \text{Fin}(C) := \{ A \subseteq C | |A| \in \mathbb{N} \} \) is the set of finite subsets of \( C \) and \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) is the restriction of the state \( h(i) \) to the space \( \text{Sp} \) and therefore the structure on that space.

\footnote{Specifically, the Riemann series theorem states that any conditionally convergent series can be reordered to have arbitrary values.} \footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]

where \( h(i)|_{\text{Sp}} : \text{Sp} \to S : c \mapsto h(i)(c) \) denotes the restriction of \( h(i) \) to \( \text{Sp} \), i.e. the structure on \( \text{Sp} \) in the state \( h(i) \). If no such relation can be established then the two histories are arguably incomparable or may be called approximately equally good. Again, the ordering could be important in some cases, see footnote.\footnote{It is very important to differentiate the series from its value. Otherwise, one may identify the series with positive or negative infinity or as being undefined. Two infinite values of the series would then not be comparable anymore, which Bostrom \cite{bostrom2011} identified as a problem for (consequentialist) ethics. But this problem can sometimes be eliminated by comparing the series itself to another. In this particular case, a history \( h \) is better than another history \( h' \), if}

\[ \sum_i \sum_{\text{Sp}} \sum_u u(h) P(u|\text{Sp} \in i) \geq u(h') P(u|\text{Sp} \in i) > 0, \]
Hare’s two-level consequentialism. (Hare, 1981, p. 25ff.) Also, most preference utilitarians ascribe preferences only to humans (or abstract agents) and do not contain prioritization among individuals, (Harsanyi, 1982, p. 46) or they use a low number of classes of moral standing. (Singer, 1993, pp. 101ff., 283f.) Whereas some have pointed out that a variety of behavior and even trivial systems can be viewed from an “intentional stance”, (Dennett, 1971; Dennett, 1989, especially pp. 29f.; compare Hofstadter, 2007, pp. 52ff.) only relatively recent articles in preference utilitarianism have discussed the connection between goal-directed behavior and ethically relevant preferences and with the universality of the former pointed out the potential universality of the latter. (Tomasik, 2015a, ch. 7; Tomasik, 2015b, ch. 4, 6; Tomasik, 2015c) This idea is an important step when formalizing preference utilitarianism because otherwise one would have to define moral standing depending on other, usually binary, notions: being alive, the ability to suffer (Bentham, 1823, ch. 17 note 122), personhood (Gruen, 2014, ch. 1), free will, sentience and (self-)consciousness (Singer, 1993, pp. 101ff.) or the ability of moral judgment. However, all of them seem to be very difficult to define (universally) in physical systems in the intended binary sense. Also, continuous definitions of these terms are often connected with goal-directed behavior. (McLaren, 2011, p. 297; Gips, 2011, p. 251) there has been some formal work at the intersection of (utilitarian) ethics, game theory and economics, most notably by Harsanyi (1982). Some formalization has also been conducted in the realm of machine ethics. (M. Anderson, S. L. Anderson, and Armen, 2004; Gips, 2011, pp. 245ff.) However, influenced by game theory and dualist traditions in philosophy, they are based on the classic agent-environment-model as displayed in figure 1 and assume utility functions (or even the utilities in different trajectories themselves) as given by the world model. Nonetheless, there is at least one parallel: all models of utilitarianism contain the notion of summing the utility over all agents. As shown in figure 3, both the definition of all agents and how to obtain the utility or welfare of an agent differ among formalizations.

In Artificial Intelligence, the idea of learning preferences has become more popular, e.g. see Fürnkranz and Hüllermeier (2010) and Nielsen and Jensen (2004) for technical treatments or Bostrom (2014, pp. 192ff.) for an introduction in the context of making an AI do what the engineers value. However, most of the time, the agent is still presumed to be separated from the environment. Nonetheless, the idea of evaluating space-time-embedded intelligence is beginning to be established in artificial (general) intelligence, (Orseau and Ring, 2012) which is closely related to the probability distribution ratio $P(str|u)/P(str)$.  

For example, Wolfram (2002, pp. 823-825, 1178-1180) and Emmeche (1997) discuss the property of life, Hofstadter (2007, pp. 9-24, 51-54) discusses consciousness and Arneson (1998, p. 5) discusses personhood.
Figure 3: Comparison between formalizations of utilitarianism. The first row shows the formalization of this paper, the second row is adapted from Harsanyi (1982, p. 46), and the third row from Gips (2011, p. 245). The utility of an agent \( n \) is denoted by \( U_n \) and its weight by \( w_n \).

5 Conclusion

Our formalization of preference utilitarianism can potentially function as a specification for an artificial intelligence or more generally as a basis for “paradise engineering” (e.g. see Ettinger, 2009, p. 124). However, there are several potential problems that require further work, before such practical applications can be approached:

- First, formal preference utilitarianism seems to differ not only from common ethical but also (preference) utilitarian intuitions. This makes specialized ethical discussion necessary. Questions of how, for example, the mechanism for interpersonal comparisons work in the presented formal framework need to be answered.

- The main imperative presented in this paper allows for some variability based on the subjectivity of prior probability distributions and the codomain of utility functions.

- Depending on the chosen probability distributions \( P(\text{str}@i|u)/P(\text{str}) \) and \( P(u) \), our formalization will usually be incomputable. This makes both discussion and application very difficult.

Acknowledgements

I am grateful to Brian Tomasik for giving me important comments that led me to systematize my formalization. I also thank Adrian Hutter for an interesting discussion on the formalization, as well as Alina Mendt, Duncan Murray, Henry Heinemann, Juliane Kraft and Nils Weller for reading and commenting on earlier versions of the paper.
References

Anderson, Michael and Susan Leigh Anderson, eds. (2011). *Machine Ethics*. Cambridge University Press. URL: [http://www.cambridge.org/cr/academic/subjects/computer-science/artificial-intelligence-and-natural-language-processing/machine-ethics](http://www.cambridge.org/cr/academic/subjects/computer-science/artificial-intelligence-and-natural-language-processing/machine-ethics).

Anderson, Michael, Susan Leigh Anderson, and Chris Armen (2004). “Towards Machine Ethics”. In: *AAAI-04 Workshop on Agent Organizations: Theory and Practice*. American Association for Artificial Intelligence. URL: [http://www.researchgate.net/publication/259656154_Towards_Machine_Ethics](http://www.researchgate.net/publication/259656154_Towards_Machine_Ethics).

Anderson, Susan Leigh (2011). “How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better”. In: *Machine Ethics*. Ed. by M. Anderson and S.L. Anderson. Cambridge University Press, pp. 524–530.

Arneson, Richard J. (1998). *What, if anything, renders all humans morally equal?* URL: [http://philosophyfaculty.ucsd.edu/faculty/rarneson/singer.pdf](http://philosophyfaculty.ucsd.edu/faculty/rarneson/singer.pdf).

Bentham, Jeremy (1823). *An Introduction to the Principles of Morals and Legislation*. Clarendon Press. URL: [http://www.econlib.org/library/Bentham/bnthPMLCover.html](http://www.econlib.org/library/Bentham/bnthPMLCover.html).

Bostrom, Nick (2011). “Infinite Ethics”. In: *Analysis and Metaphysics* 10, pp. 9–59.

— (2014). *Superintelligence. Paths, Dangers, Strategies*. 1st ed. Oxford University Press.

Dennett, Daniel (1971). “Intentional Systems”. In: *The Journal of Philosophy* 68.4, pp. 87–106. URL: [http://www.jstor.org/stable/2025382](http://www.jstor.org/stable/2025382).

— (1989). *The Intentional Stance*. MIT Press.

— (2006). *Computers as Prostheses for the Imagination*. Talk at The International Computers and Philosophy Conference. Laval, France.

Downey, Allen B. (2012). *Think Complexity*. O’Reilly.

Emmeche, Claus (1997). Center for the Philosophy of Nature and Science Studies Niels Bohr Institute. URL: [http://www.nbi.dk/~emmeche/cePubl/97edefLife.v3f.html](http://www.nbi.dk/~emmeche/cePubl/97edefLife.v3f.html).

Ettinger, Robert C. W. (2009). *Youniverse: Toward a Self-Centered Philosophy of Immortalism and Cryonics*. URL: [http://books.google.de/books/about/Youniverse.html?id=xer8lRSjKSwC&redir_esc=y](http://books.google.de/books/about/Youniverse.html?id=xer8lRSjKSwC&redir_esc=y).

Fünkranz, Johannes and Eyke Hüllmeier, eds. (2010). *Preference Learning*. Springer.

Gips, J. (2011). “Towards the Ethical Robot”. In: *Machine Ethics*. Ed. by M. Anderson and S.L. Anderson. 1st ed. Cambridge University Press, pp. 244–253.

Gruen, Lori (2014). ed. Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: [http://plato.stanford.edu/archives/fall2014/entries/moral-animal](http://plato.stanford.edu/archives/fall2014/entries/moral-animal).

Hare, Richard Mervyn (1981). *Moral thinking. Its levels method, and point*. Two-level consequentialism. Clarendon Press.
Harsanyi, John C. (1982). “Morality and the theory of rational behaviour”. In: Utilitarianism and Beyond. Ed. by Amartya Sen and Bernard Williams. Cambridge University Press. Chap. 2, pp. 39–62. URL: http://ebooks.cambridge.org/chapter.jsf?bid=CB09780511611964&cid=CB09780511611964A009&tabName=Chapter
Hawking, Stephen and Leonard Mlodinow (2010). The Grand Design. Bantam B.
Hofstadter, Douglas (2007). I Am a Strange Loop. Basic Books.
Legg, Shane (1997). “Solomonoff Induction”. MA thesis. University of Auckland. URL: http://www.vetta.org/documents/legg-1996-solomonoff-induction.pdf
McLaren (2011). “Computation Models of Ethical Reasoning. Challenges, Initial Steps, and Future Directions”. In: Machine Ethics. Ed. by M. Anderson and S.L. Anderson. 1st ed. Cambridge University Press, pp. 297–315.
Moor, James H. (2011). “The Nature, Importance, and Difficulty of Machine Ethics”. In: Machine Ethics. Ed. by Michael Anderson and Susan Leigh Anderson. Cambridge University Press. Chap. 1, pp. 13–20.
Nielsen, Thomas D. and Finn V. Jensen (2004). “Learning a decision maker’s utility function from (possibly) inconsistent behavior”. In: Artificial Intelligence 160.1-2, pp. 53–78. URL: http://www.sciencedirect.com/science/article/pii/S0004370204001328
Nozick, Robert (1969). “Newcomb’s Problem and Two Principles of Choice”. In: Essays in Honor of Carl G. Hempel. Ed. by Nicholas Rescher et al. Springer, pp. 114–146. URL: http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf
Olshausen, Bruno A. (2004). Bayesian probability theory.
Orseau, Laurent and Mark Ring (2012). “Space-Time Embedded Intelligence”. In: Artificial General Intelligence. Ed. by Joscha Bach, Ben Goertzel, and Matthew Iklé. Vol. 5, p. 391. URL: http://agi-conference.org/2012/wp-content/uploads/2012/12/paper_76.pdf
Robert, C.P. (1994). The Bayesian Choice. A Decision-Theoretic Motivation. 1st ed.
Schmidhuber, Jürgen (1999). A Computer Scientist’s View of Life, the Universe, and Everything. URL: http://arxiv.org/pdf/quant-ph/9904050v1.pdf
Shiffman, Daniel (2012). The Nature of Code. Ed. by Shannon Fry. URL: http://natureofcode.com/book/
Singer, Peter (1993). Practical Ethics. 2nd ed. Cambridge University Press.
Tomasik, Brian (2015a). Do Video-Game Characters Matter Morally? URL: http://www.webcitation.org/6X7w4AJnb
— (2015b). Hedonistic vs. Preference Utilitarianism. URL: http://www.webcitation.org/6X7vepyJP
— (2015c). Is There Suffering in Fundamental Physics? URL: http://www.webcitation.org/6X7vte3XP
Wolfram, Stephen (1983). “Cellular Automata”. In: Los Alamos Science 9, pp. 2–21. URL: http://www.stephenwolfram.com/publications/academic/cellular-automata.pdf
Wolfram, Stephen (2002). *A New Kind of Science*. Wolfram Media. URL: [http://www.wolframscience.com/nksonline/toc.html](http://www.wolframscience.com/nksonline/toc.html).

Yudkowsky, Eliezer (2001). *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Machine Intelligence Research Institute. URL: [http://intelligence.org/files/CFAI.pdf](http://intelligence.org/files/CFAI.pdf).

Zuse, Konrad (1967). “Rechnender Raum”. In: *Elektronische Datenverarbeitung* 8, pp. 336–344. URL: [ftp://ftp.idsia.ch/pub/juergen/zuse67scan.pdf](ftp://ftp.idsia.ch/pub/juergen/zuse67scan.pdf).

— (1969). *Rechnender Raum*. Vol. 1. Vieweg & Sohn.