Synthese
DOI 10.1007/s11229-015-0883-1

Formalizing preference utilitarianism in physical world models

Caspar Oesterheld[1]

Received: 23 April 2015 / Accepted: 27 August 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 is a first step in providing 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 Anderson and 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 are assumed to be ontologically fundamental.

✉ Caspar Oesterheld
caspar.oesterheld@uni-bremen.de

[1] University of Bremen, Bremen, Germany

Published online: 16 September 2015

In this paper however, we take the following steps towards a workable and simple formalization of preference utilitarianism[1] in physical world models:

- We describe the problem of informality in ethics and the shortcomings of previous dualist approaches to formalizing utilitarian ethics (Sect. 2).
- 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 (Sect. 3).
- We compare our approach with existing work in ethics, game theory and artificial intelligence (Sect. 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; 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: 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 Fig. 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. 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: (compare Orseau and Ring 2012)

- 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

[1] For introductions to and ethical discussions of the underlying notion of preference utilitarianism see Tomasik (2015a, b).

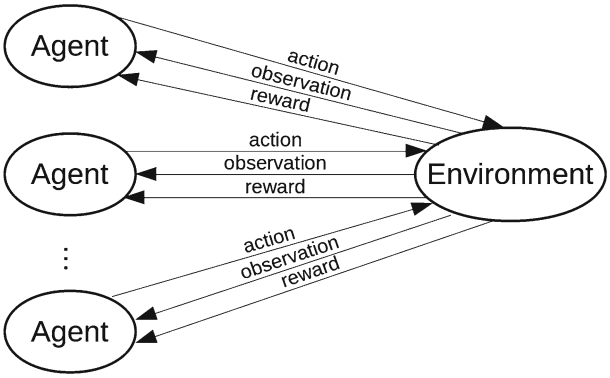

Fig. 1 The classic agent-environment-model

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, (Wolfram 2002, ch. 9; Schmidhuber 1999; Zuse 1967, 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 Sect. 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.

Fig. 2 A state of a two-dimensional cellular automaton. It is very unclear, what agents are and which preferences they have. Adapted from http://en.wikipedia.org/wiki/Conway%27s_Game_of_Life#mediaviewer/File:Conways_game_of_life_breeder

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 $d : S^C \to S^C$ that maps a *world state* $s : C \to S$ 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.[2]

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 $str : Spc \to S$ that maps the cells of the space onto cell values.

[2] 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.

A *history* is a function $h : \mathbb{N} \to S^C$ 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

Before extracting preferences from a given structure, we have to decide on a model of preferences. Preferences themselves are mere orderings of alternative outcomes or lotteries over these outcomes with the outcomes being entire histories $h \in (S^C)^{\mathbb{N}}$ in our case. The problem is that this makes it difficult to compare two outcomes when the preferences of multiple individuals are involved. To be able to make such comparisons, we move from orderings to utility functions $u : (S^C)^{\mathbb{N}} \to \mathbb{R}$ that map histories of the world onto their (cardinal) utilities.[3] This will make it possible to just add up the utilities of different individuals and then compare the sum among outcomes. This by no means "solves" the problem of interpersonal comparison. Rather, it makes it more explicit. For example, a given set of preferences is represented equally well by u and $2 \cdot u$, but *ceteris paribus* $2 \cdot u$ will make the preferences more significant in summation. Different approaches to the problem have been proposed. (Hammond 1989) In this paper we will ignore the problem (or hope that the fair treatment in determining all individuals' utility functions induces moral permissibility). Now we ask the question: Does a particular structure str want to maximize some utility function u?

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.[4]

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.[5] Here, the utility function is

[3] Other codomains of utility functions seem possible as long as they are subsets of a totally ordered vector space over $\mathbb{R}$. Intervals like $[0, 1]$ seem specifically suitable, because they avoid problems of infinite utility and allow for normalization. (Isbell 1959)

[4] 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.

[5] 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

interpreted as a hypothesis about the structure's "true intentions"[6]. 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 utility function u is the goal of structure *str*.
- Maximizing u was the 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

$$P(u|str@i) = \frac{P(str@i|u) \cdot P(u)}{P(str@i)}, \tag{1}$$

where $P(u)$ and $P(str@i)$ 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 $P(str@i)$ 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.

Footnote 5 continued
history, because then structures cannot have preferences 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).

[6] Intuitively, some structures can have more than one utility function, while others have no utility function at all. One way to model this would be to understand different utility functions as events in separate sample spaces. So, the sum $\sum_u P(u|str)$ could vary among different structures str. A similar scenario is the inference of multiple diseases from a set of symptoms. (Charniak 1983) While some individuals may have no diseases or preferences at all, others may be thought of as having more than one disease or utility function. In more technical terms, for each utility function there would be a sample space of having that utility function and not having that utility function.
In this paper however, we will assume mutual exclusivity and collective exhaustiveness of utility functions. All utility functions live in the same sample space and thus $\sum_u P(u|str) = 1$ for all structures str. This does not mean that all structures have equal moral standing: The idea is that "meaningless" structures str have high $P(u|str)$ only for constant utility functions u, i.e. for "don't care"-utility functions, which are irrelevant for decision making.

Nonetheless, there seem to be canonical approaches. $P(str@i|u)$ should be understood as the probability that str is chosen at time step i by an approximately rational agent that wants to maximize u. So, structures that are better at maximizing or more suitable for u should receive higher $P(str@i|u)$ values. This corresponds to the assumption of (approximate) rationality in Dennett's intentional stance. (Dennett 1989, pp. 21, 49f.) Unfortunately, the debate about causal and evidential decision theory (e.g. Peterson 2009, ch. 9) shows that formalizing the notion of rational choice is difficult.

The prior of utility functions $P(u)$ on the other hand should denote the "intrinsic plausibility" of a goal u. That does not have to mean defining and excluding "evil" or "banal" utility functions. In the preference extraction context, utility functions are models or hypotheses that explain the behavior of a structure. And Solomonoff's formalization of Occam's razor is often cited as a universal prior distribution of hypotheses. (Legg 1997) It assumes complicated hypotheses (utility functions), i.e. ones that require more symbols to be described in some programming language, to be less likely than simpler ones.

If utility functions are conceived of as competing hypotheses (see footnote 6), then

$$P(str@i) = \sum_u P(str@i|u) \cdot P(u).$$

Otherwise, $P(str@i)$ could potentially be chosen more freely.

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 function 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 on a given space and time step 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

$$U_{str@i} = \sum_u P(u|str@i)u(h), \tag{2}$$

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

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 Eq. 2 and rely on (intuitively) irrelevant ones to receive high $P(u|str@i)$ values only for constant and therefore irrelevant utility functions u (see footnote 6). To sum the utility over all agents, we not only have to sum over all structures in a particular state, but 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 Eq. 2

$$\sum_i \sum_{str@i} U_{str@i} = \sum_i \sum_{str@i} \sum_u P(u|str@i)u(h), \tag{3}$$

where $U_{str@i}$ denotes the welfare or utility of the structure 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.[8] So, our formalization of the main imperative of preference utilitarianism turns out to be nothing more than maximizing *global, all-time expected utility* (of every space-time embedded agent that ever comes into existence). In general, the value of the series depends on the order of

[7] If Solomonoff's prior is chosen for $P(u)$, all incomputable utility functions have zero probability. Since the set of computable functions is countable, only countably many summands could possibly be non-zero.

[8] More precisely, but less elegantly, one could write

$$\sum_{i=0}^{\infty} \sum_{Spc \in Fin(C)} \sum_{u:(S^C)^{\mathbb{N}} \to \mathbb{R}} u(h) P(u|(h(i)|_{Spc})@i),$$

where $Fin(C) := \{A \subseteq C || A| \in \mathbb{N}\}$ is the set of finite subsets of C and $h(i)|_{Spc} : Spc \to S : c \mapsto h(i)(c)$ is the restriction of the state $h(i)$ to the space Spc and therefore the structure on that space.

these infinite sums.[9] Also, the series can diverge. Nonetheless it may still be usable for comparing histories in many cases.[10]

4 Related work

Preferentist utilitarianism has become a common form of utilitarianism in the second half of the 20th century, with the best known proponents being Hare and Singer. However, the intuitions underlying the presented formalization are different from the most common ethical intuitions in preference utilitarianism. Since our formal preference utilitarianism is not meant to describe a decision procedure for humans (or, more generally and in Hare's (1981, pp. 44f.) terminology, non-"archangels"), we do not consider an application-oriented utilitarianism like 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, 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 2015b, ch. 7; Tomasik 2015a, 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.[11] Also, continuous definitions of these terms are often connected with goal-directed behavior. (Tomasik 2015a, ch. 4; Wolfram 2002, p. 1136)

[9] Specifically, the Riemann series theorem states that any conditionally convergent series can be reordered to have arbitrary values.

[10] 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 (2011) 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

$$\sum_i \sum_{Spc} \sum_u u(h) P(u|(h(i)|_{Spc})@i) - u(h') P(u|(h'(i)|_{Spc})@i) > 0,$$

where $h(i)|_{Spc} : Spc \to S : c \mapsto h(i)(c)$ denotes the restriction of $h(i)$ to Spc, i.e. the structure on Spc 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 9.

[11] 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.

sum over all agents	utility of an agent
$\sum_{i} \sum_{str@i}$	$\sum_{u} P(u\|str@i)u(h)$
$\sum_{\text{agent } n \in N}$	U_n
$\sum_{\text{agent } n \in N}$	$w_n \cdot U_n$

Fig. 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

Whereas most ethical work is conducted informally, (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. (Anderson et al. 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 Fig. 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 Fig. 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 $P(str@i|u)$.

5 Conclusion

By reversing Dennett's intentional stance with Bayes' theorem, we were able to ascribe preferences to physical objects and thus formalize preference utilitarianism in cellular automata. Theoretically, such formalizations can 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 of our formalization or improved variations of it can be approached:

– Through sums over all structures, possible utility functions and states and the application of incomputable concepts like Solomonoff's prior in $P(u)$, our formalization is incomputable in theory and practice. So even in simulations of cellular automata our formalization is not immediately applicable.

- Computing our global welfare function in the real world is even more difficult, because it requires full information about the world on particle level. Also, the formalization must first be translated into the physical laws of our universe.
- The difficulty to apply our formalization is by no means only relevant to actually using it as a moral imperative. Instead, it is also relevant to discussing our formalization from a normative standpoint: Even though the derivation of our formalization is plausible, it may still differ significantly from intuition. There could be some kind of trivial agents with trivial preferences that dominate comparison of different histories. Because the formalization's incomputability makes it difficult to assess whether such problems are present, further work on its potential flaws is necessary. Based on such discussion, our formalization may be revised or even discarded. In any case, we could learn a lot from its shortcomings especially due to the formalization's simplicity and plausible derivation.
- We outlined how $P(str@i|u)$ and $P(u)$ could be determined in principle. However, they need to be specified more formally, which in the case of $P(str@i|u)$ seems to require a solution to the problem of normative decision theory. Some problems of our formalization could inspire additional refinements of these distributions.

Acknowledgments 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. I owe thanks to the two anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

References

Anderson, S. L. (2011). How machines might help us achieve breakthroughs in ethical theory and inspire us to behave better. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 524–530). Cambridge: Cambridge University Press.

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge: Cambridge University Press. http://www.cambridge.org/cr/academic/subjects/computer-science/artificial-intelligence-and-natural-language-processing/machine-ethics.

Anderson, M., Anderson, S. L. & Armen, C. (2004). Towards machine ethics. In: *AAAI-04 Workshop on agent organizations: Theory and practice*. American Association for Artificial Intelligence, Menlo Park. http://www.researchgate.net/publication/259656154_Towards_Machine_Ethics.

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

Bentham, J. (1823). *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press. http://www.econlib.org/library/Bentham/bnthPMLCover.html.

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics*, *10*, 9–59.

Bostrom, N. (2014). *Superintelligence. Paths, dangers, strategies* (1st ed.). Oxford: Oxford University Press.

Charniak, E. (1983). The Bayesian basis of common sense medical diagnosis. In: *AAAI-83 Proceedings*, Washington, DC.

Dennett, D. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106. http://www.jstor.org/stable/2025382.

Dennett, D. (1989). *The intentional stance*. Cambridge: mit press.

Dennett, D. (2006). *Computers as prostheses for the imagination*. In: *Talk at The International Computers and Philosophy Conference.* Laval, France.
Downey, A. B. (2012). *Think complexity*. Sebastopol: O'Reilly.
Emmeche, C. (1997). Center for the Philosophy of Nature and Science Studies Niels Bohr Institute. http://www.nbi.dk/~emmeche/cePubl/97e.defLife.v3f.html.
Ettinger, R. C. W. (2009). *Youniverse: Toward a self-centered philosophy of immortalism and cryonics*. Boca Raton: Universal-Publishers
Fürnkranz, J., & Hüllermeier, E. (Eds.). (2010). *Preference learning*. Berlin: Springer.
Gips, J. (2011). Towards the ethical robot. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (1st ed., pp. 244–253). Cambridge: Cambridge University Press.
Gruen, L. (2014). In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2014/entries/moral-animal.
Hammond, P. J. (1989). *Interpersonal comparison of utility: Why and how they are and should be made*. http://homepages.warwick.ac.uk/~ecsgaj/icuSurvey.pdf.
Hare, R. M. (1981). *Moral thinking. Its levels method, and point. Two-level consequentialism*. Oxford: Clarendon Press.
Harsanyi, J. C. (1982). Morality and the theory of rational behaviour. In A. Sen & B. Williams (Eds.), *Utilitarianism and beyond, Chap. 2* (pp. 39–62). Cambridge: Cambridge University Press. http://ebooks.cambridge.org/chapter.jsf?bid=CBO9780511611964&cid=CBO9780511611964A009&tabName=Chapter.
Hawking, S. & Mlodinow, L. (2010). *The grand design*. New York: Bantam B.
Hofstadter, D. (2007). *I am a strange loop*. New York: Basic Books.
Isbell, J. R. (1959). Absolute games. In A. W. Tucker & R. D. Luce (Eds.), *Contributions to the theory of games* (Vol. 4, pp. 357–396). Princeton: Princeton University Press.
Legg, S. (1997). Solomonoff induction. MA Thesis. University of Auckland. http://www.vetta.org/documents/legg-1996-solomonoff-induction.pdf.
McLaren, B. N. (2011). Computation models of ethical reasoning. Challenges, initial steps, and future directions. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (1st ed., pp. 297–315). Cambridge: cambridge University Press.
Moor, J. H. (2011). The nature, importance, and difficulty of machine ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics, Chap. 1* (pp. 13–20). Cambridge: Cambridge University Press.
Nielsen, T. D. & Jensen, F. V. (2004). Learning a decision maker's utility function from (possibly) inconsistent behavior. In: *Artificial Intelligence*, *160*(1–2), 53–78. http://www.sciencedirect.com/science/article/pii/S0004370204001328.
Nozick, R. (1969). Newcomb's problem and two principles of choice. In: N. Rescher et al. (Ed.), *Essays in honor of Carl G. Hempel* (pp. 114–146). Berlin: Springer. http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf.
Olshausen, B. A. (2004). *Bayesian probability theory*. Retrieved from http://redwood.berkeley.edu/bruno/npb163/bayes.pdf
Orseau, L. & Ring, M. (2012). Space-time embedded intelligence. In: J. Bach, B. Goertzel & M. Iklé (Eds.), *Artificial general intelligence, Vol. 5* (p. 391). http://agi-conference.org/2012/wp-content/uploads/2012/12/paper_76.pdf.
Peterson, M. (2009). *An introduction to decision theory*. Cambridge: Cambridge University Press.
Robert, C. P. (1994). *The Bayesian choice. A decision-theoretic motivation* (1st ed.). Berlin: Springer.
Schmidhuber, J. (1999). *A computer scientist's view of life, the universe, and everything*. http://arxiv.org/pdf/quant-ph/9904050v1.pdf.
Shiffman, D. (2012). In S. Fry (Ed.), *The nature of code*. http://natureofcode.com/book/.
Singer, P. (1993). *Practical ethics* (2nd ed.). Cambridge: Cambridge University Press.
Tomasik, B. (2015a). *Do video-game characters matter morally?* http://www.webcitation.org/6X7w4AJnb.
Tomasik, B. (2015b). *Hedonistic vs. preference utilitarianism*. http://www.webcitation.org/6X7vepyJP.
Tomasik, B. (2015c). *Is there suffering in fundamental physics?* http://www.webcitation.org/6X7vte3XP.
Wolfram, S. (1983). Cellular automata. In: *Los Alamos Science*, *9*, 2–21. http://www.stephenwolfram.com/publications/academic/cellular-automata.pdf.
Wolfram, S. (2002). *A new kind of science*. Champaign: Wolfram Media. http://www.wolframscience.com/nksonline/toc.html.

Yudkowsky, E. (2001). *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. Machine Intelligence Research Institute. http://intelligence.org/files/CFAI.pdf.

Zuse, K. (1967). Rechnender Raum. In: *Elektronische Datenverarbeitung*, *8*, 336–344. ftp://ftp.idsia.ch/pub/juergen/zuse67scan.pdf.

Zuse, K. (1969). *Rechnender Raum*. Brunswick: Vieweg & Sohn.