 doi:10.20944/preprints201805.0385.v1

Article

Towards a Formalization of Hedonic Well-being in Physicalist World Models

Caspar Oesterheld[1,†]

[1] Foundational Research Institute; caspar.oesterheld@foundational-research.org

† Current address: Hardenbergstr. 9, 10623 Berlin, Germany.

Abstract: Formally specifying hedonic well-being is difficult, but relevant to utilitarian theories of morality. In this paper we describe a starting point based on attitudinal hedonism, which posits that hedonic well-being is determined by the extent to which a moral patient believes her preferences to be fulfilled. While this assumption probably does not capture the sought-out notion of well-being fully, our formalization seems to be relevant first step towards more satisfactory definition of hedonic well-being.

Keywords: hedonic well-being; utilitarianism; machine ethics; attitudinal hedonism; subjective desire satisfactionism

1. Introduction

Recently, growing effort has been made to formalize ethics and various ethically relevant concepts, not only to be more precise and avoid misunderstandings but also to be able to make machines behave ethically [1, p. 297; 2, p. 251; 3, p. 19; 4]. [5] describes a formalization of preference utilitarianism in world models in which agents are not ontologically fundamental entities but physical objects, in particular structures in a cellular automaton. In this paper we point out how this formalization can be modified to obtain a formalization of a specific version of hedonistic utilitarianism, in which happiness is equated with believing that one's preferences are fulfilled.

2. A brief recap of the main ideas behind formal preference utilitarianism

Oesterheld's [5] formalization of preference utilitarianism defines the subjective well-being of a space-time embedded agent a at time step i as its expected utility over different utility functions

$$U_{a@i} = \sum_u P(u \mid a@i) u(h), \tag{1}$$

where h is the true trajectory or outcome of the system a is living in, $P(u \mid a@i)$ is the probability that a's goal is to maximize u and $u(h)$ is the utility of h to a if a's utility function were u. The preference extraction $P(u \mid a@i)$ is based on Bayes' theorem

$$P(u \mid a@i) = \frac{P(a@i \mid u) P(u)}{P(a@i)}, \tag{2}$$

where the main component $P(a@i \mid u)$ is to be understood as testing how suitable a is for maximizing u in the spirit of functionalism, behaviorism and Dennett's [6] intentional stance [5, ch. 3.2]. From the utility $U_{a@i}$ of a single agent a, the overall utility of the trajectory h can be obtained by summing the utility of all agents that ever come into existence.

3. From extracting preferences to extracting subjective well-being

We would now like to move from a preferentist to a hedonistic conception of well-being.[1] Specifically, the welfare of an agent shall not be determined by the fulfillment of its preferences ($u(h)$ in equation 1), but by the extent to which the agent *believes* its preferences to be fulfilled. This form of hedonism is perhaps best known as attitudinal hedonism or subjective desire satisfactionism, see section 5.1 for references.

Our strategy for ascribing beliefs is grounded in the following intuition: an intelligent agent a (as containing its state of mind) contains the belief that the history has some property p (that a cares about) if a's physical configuration is usually brought about by a (or rather its predecessor(s)) "learning" in some way that p is true.[2] Therefore, if a (as containing its current state of mind) usually occurs in histories in which p is true, a should and probably will believe p to be true rather than false even if it is actually false.[3]

To formalize this strategy, we will consider what information a encodes, i.e., what one could potentially infer from a's physical configuration. This can be formalized as the expected value of the utility function u for a, i.e. $\mathbb{E}\left[u(h) \mid a\right]$, where a is not assumed to necessarily know the time step of its existence i. This gives us a new definition of an agent's welfare, its *hedonic welfare*:

$$H_{a@i} = \sum_u P(u \mid a@i)\mathbb{E}\left[u(h) \mid a\right]. \tag{3}$$

As an example, consider the agent Anna. Her primary life goal it is to preserve the environment, so – assuming the approach of [5] works – $P(u \mid \text{Anna@}i)$ is high only for utility functions u that express some aspect of how Anna cares about environmental preservation. Throughout her life, Anna has read articles and books about the state of the environment, has considered statistics, visited conservation areas and the Arctic, and has talked to many experts and politicians. Her memories (and thus her brain) encode aspects of these experiences. Thus, $\mathbb{E}\left[u(h) \mid \text{Anna}\right]$ represents the state of the environment as judged based on Anna's memories. Assuming that this value is low, our formalization would judge Anna as unhappy.

4. Discussion

We now discuss some of the issues one may have with the given formalization of attitudinal hedonism. This discussion is not intended to be exhaustive. In particular we will only discuss potential problems with our formalization rather than the target of our formalization (attitudinal hedonism) or the aspects of the formalization borrowed from [5].

Ascribing the belief $\mathbb{E}\left[u(h) \mid a\right]$ fails to capture the fact that an agent usually cannot answer queries about its entire physical configuration, let alone optimally deduce probability distributions from it. Thus, it might be possible to deduce high u values from the physical structure of a with near certainty, even if a does not "actually know" that u is fulfilled. While the outside-view assumption of rationality works well enough for preference extraction to evade immediate and outright refutation [5], it seems to fail when it comes to extracting beliefs.

A more plausible approach for extracting beliefs, and perhaps also extracting preferences, may therefore involve abstracting away from all the irrelevant details of an organism (compare, e.g., [12, ch. 3, 14]) and consider merely the algorithm the organism is implementing (cf. 13). Specifically, we

---

1 For some discussion of the differences between hedonism and preferentism, see, e.g., 7, ch. 2-5; 8, ch. 8, 9; 9, ch. 5, 6; or 10.

2 Like [5], this resembles Dennett's intentional stance, which asks us to model an agent's beliefs as what that agent ought to believe (e.g. 6, p. 17).

3 This is, of course, not the only strategy for inferring beliefs from an agent. An alternative is based on the view that beliefs are revealed by behavior [11, ch. 3.3.2]. E.g., an agent holds the belief that some box contains a desirable object if the agent attempts to open the box. A comparison of such different strategies is beyond the scope of this paper.

 doi:10.20944/preprints201805.0385.v1

could base our view on physical processes on the computational theory of mind and computational functionalism specifically. As [14, ch. 2] describes,

> [c]omputational functionalism is the view that mental states and events – pains, beliefs, desires, thoughts and so forth – are computational states of the brain [...]. The nature of the mind is independent of the physical making of the brain [...]. What matters is our functional organization: the way in which mental states are causally related to each other, to sensory inputs and to motor outputs.

Unfortunately, interpreting a physical system as a mind or as performing computation is difficult (see [15] and references therein).

A related issue is that an agent may use the information that is available to him inaccurately. In other words, the agent may hold beliefs that are not justified by the information that is available to him. For example, assume that Bob knows the base rate for car accidents per kilometer traveled in the region in which he lives. Let us say this base rate (together with knowledge of how much Bob travels) yields a probability of 0.01% that Bob will die in a car crash in the next year. However, like most people, Bob is susceptible to base rate neglect [16] and to believing that he is a better-than-average driver [17]. This leads him to feel much safer than he has reason to believe – when asked about whether he expects to die in the next year, he reports a probability of only 0.0001%. Assuming that Bob prefers not to die in a car crash, Bob is happier than equation 3 makes him out to be.

Another issue is that our formalization gives equal weight to long-held beliefs as it does to recently formed ones. Arguably, this conflicts with our subjective experience. For example, learning that one has won the lottery should increase happiness greatly in the short term, but knowing about one's wealth increases happiness only slightly in the long term. So, according to this *hedonic treadmill* account of well-being [18], plausible versions of attitudinal hedonism should focus on recent changes in beliefs [19]. While this is a promising approach in general (also see section 5.2), it seems to require us to solve the problem of personal identity to compare the beliefs of two versions of the same agent.

Alternatively, we could try to give more weight to beliefs that the agent is currently contemplating. Then happiness (suffering) would mean performing computations related to (un)fulfilled preferences. Perhaps the hedonic treadmill then arises from the fact that people mostly consider new information.

5. Related work

The main contribution of this paper is to present a formal definition of hedonic well-being that can be used for the utilitarian calculus as described by [5, ch. 3.4]. A variety of (mostly informal) conceptions of hedonic well-being have been discussed in the vast literature on the topic. In the following we describe how the presented formalization relates to informal definitions of well-being that have been discussed in the context of utilitarianism (section 5.1) and to other formal approaches (section 5.2).

5.1. Informal treatments of the presented view of well-being

In hedonistic utilitarianism, happiness is usually defined as raw pleasure (and the absence of raw pain) – whatever that is [20] – rather than as related to preferences (cf. 21 for a neuroscientific perspective). Nevertheless, defining happiness as the state of believing that one's preferences are fulfilled, is by no means a novel proposal. Instead, it has been proposed in the literature multiple times under various names:

- desire theory with knowledge-oriented modification b [7, sect. 5.2.3];
- (intrinsic) attitudinal hedonism [22; 23, ch. 3; 24, p. 9], especially in its confidence-adjusted form [25];
- the experience requirement in preference (or desire) theory [26, pp. 75-80];
- subjective desire satisfactionism [23, ch. 2; 27, ch. 4.3; cf. 28, sect. 1.2].

 doi:10.20944/preprints201805.0385.v1

Also consider the classification system described by [29], in which the presented formalization would be classified as type 1, because it is based on combining the desire and the experience requirement.

The presented formalization stands in contrast not only to raw-pleasure hedonism, but also to definitions based solely on desires regarding the mental state, or "experience-directed desires" [30, ch. 3.4], such as that of Henry Sidgwick [8, ch. 9], Lukas Gloor [31], Parfits's [32] preference-hedonism [33, sect. 2] or the experience-oriented success theory (see, e.g., 7, sect. 5.2.1).[4]

5.2. Other formal definitions of subjective well-being

[19] formalizes hedonic well-being (happiness) for reinforcement learners based on how incoming rewards and observations differ from expectation. For example, if an agent expects to receive very small rewards in the future, it will be happy if it receives slightly larger (but still small) rewards. Similarly, observations that lead to adjusting expectations of future rewards upward contribute to happiness. This idea is based on experimental results in psychology [34] including the hedonic treadmill account of well-being on the one hand, and the temporal difference technique in reinforcement learning (see 35, ch. 6) on the other hand. It also operates on a relatively high-level, i.e. non-physical reinforcement learners, which already have a reward system in place, as opposed to dealing with the pure physics of our approach.

6. Conclusion

Based on common versions of hedonism, we described what is – to my knowledge – the first general formalization of hedonic well-being. Unfortunately, the formalization differs significantly from intuition in a number of ways. For each of these we proposed ideas for modifications that might remedy such problems. We thus hope that the present formalization lays the groundwork for further research in formalizing hedonistic accounts of well-being.

Funding: This research received no external funding.

Acknowledgments: I thank Brian Tomasik, Lukas Gloor, David Althaus and Max Daniel for helpful comments and discussion. I am also grateful to Simon Knutsson and Kyle Bogosian for valuable references.

Conflicts of Interest: The authors declare no conflict of interest.

1. McLaren, B. Computation Models of Ethical Reasoning. Challenges, Initial Steps, and Future Directions. In *Machine Ethics*, 1 ed.; Cambridge University Press: Cambridge, 2011; pp. 297–315.
2. Gips, J. Towards the Ethical Robot. In *Machine Ethics*, 1 ed.; Cambridge University Press: Cambridge, 2011; pp. 244–253.
3. Moor, J.H. The Nature, Importance, and Difficulty of Machine Ethics. In *Machine Ethics*; Cambridge University Press, 2011; chapter 1, pp. 13–20.
4. Muehlhauser, L.; Helm, L. Intelligence Explosion and Machine Ethics. In Singularity Hypotheses: A Scientific and Philosophical Assessment, 2012.
5. Oesterheld, C. Formalizing Preference Utilitarianism in Physical World Models. *Synthese* **2016**, *193*, 2747–2759. doi:10.1007/s11229-015-0883-1.
6. Dennett, D. *The Intentional Stance*; MIT Press, 1989.
7. Brülde, B. The Human Good. PhD thesis, University of Gothenburg, Gothenburg, 1998.
8. de Lazari-Radek, K.; Singer, P. *The Point of View of the Universe. Sidgwick & Contemporary Ethics*; Oxford University Press, 2014.
9. Tännsjö, T. *Hedonistic Utilitarianism*; Edinburgh University Press, 1998.

[4] The literature does not always make a clear distinction between wellbeing as having one's desires about one's state of consciousness fulfilled versus holding (conscious) beliefs about one's desires (about the world, not only one's mind) being fulfilled.

10. Tomasik, B. Hedonistic vs. Preference Utilitarianism. https://foundational-research.org/publications/hedonistic-vs-preference-utilitarianism/.
11. Hájek, A. Interpretations of Probability. In *The Stanford Encyclopedia of Philosophy*, Winter 2012 ed.; Zalta, E.N., Ed.; 2012.
12. Hofstadter, D. *I Am a Strange Loop*; Basic Books, 2007.
13. Francescotti, R. Supervenience and Mind. Internet Encyclopedia of Philosophy.
14. Shagrir, O. *The Rise and Fall of Computational Functionalism*; Contemporary Philosophy in Focus, Cambridge University Press, 2005; chapter 9, pp. 220–250.
15. Oesterheld, C. Bayesian Decision Theory, Physicalism and the Problem of Building Phenomenological Bridges. Unpublished manuscript.
16. Tversky, A.; Kahneman, D. Evidential Impact of Base Rates. In *Judgment under uncertainty: Heuristics and biases*; Cambridge University Press, 1982; chapter 10, pp. 153–160.
17. McCormick, I.A.; Walkey, F.H.; Green, D.E. Comparative Perceptions of Driver Ability – A Confirmation and Expansion. *Accident Analysis and Prevention* **1986**, *18*, 205–208.
18. Frederick, S. Hedonic Treadmill. In *Encyclopedia of Social Psychology*; SAGE, 2007; pp. 419–420.
19. Daswani, M.; Leike, J. A Definition of Happiness for Reinforcement Learning Agents. Artificial General Intelligence, 8th International Conference. Springer, 2015, Vol. 9205, *Lecture Notes in Computer Science*, pp. 231–240.
20. Weijers, D. Hedonism. In *Internet Encyclopedia of Philosophy*.
21. Berridge, K.C.; Robinson, T.E.; Aldridge, J.W. Dissecting Components of Reward: 'Liking', 'Wanting', and Learning. *Current Opinion in Pharmacology* **2009**, *9*, 65–73.
22. Feldman, F. The Good Life: A Defense of Attitudinal Hedonism. In *Ethical Theory. An Anthology*, 2 ed.; Wiley-Blackwell, 2013; Vol. 34, *Blackwell Philosophy Anthologies*, chapter 31, pp. 266–276.
23. Heathwood, C. Desire Satisfactionism and Hedonism. *Philosophical Studies* **2006**, *128*, 539–563. doi:10.1007/s11098-004-7817-y.
24. Brülde, B. Happiness and the Good Life. Introduction and Conceptual Framework. *Journal of Happiness Studies* **2007**, *8*, 1–14. doi:10.1007/s10902-006-9002-9.
25. Søraker, J.H. Introducing Confidence – Adjusted Intrinsic Attitudinal Hedonism (CAIAH), and its Implications for Ethics of Technology. SPT 2013: Technology in the Age of Information, 18th International Conference of the Society for Philosophy and Technology; , 2013; pp. 244–246.
26. Mulgan, T. *Understanding Utilitarianism*; Understanding Movements in Modern Thought, Routledge, 2014.
27. Heathwood, C. Desire-Satisfaction Theories of Welfare. PhD thesis, Graduate School of the University of Massachusetts Amherst, 2005.
28. Moen, O.M. An Argument for Hedonism. *Journal of Value Inquiry* **2016**, *50*, 267–281.
29. Woodard, C. Classifying Theories of Welfare. *Philosophical Studies* **2013**, *165*, 787–803. doi:10.1007/s11098-012-9978-4.
30. Bain, D.; Brady, M. Pain, Pleasure, and Unpleasure. *Review of Philosophy and Psychology* **2014**, *5*, 1–14. doi:10.1007/s13164-014-0176-5.
31. Gloor, L. Tranquilism. https://foundational-research.org/tranquilism/.
32. Parfit, D. What Makes Someone's Life Go Best. In *Ethical Theory. An Anthology*, 2 ed.; Wiley-Blackwell, 2013; Vol. 34, *Blackwell Philosophy Anthologies*, chapter 34, pp. 294–298. reprinted from Reasons and Persons (Oxford University Press, 1984), pp. 493–502.
33. Rønnow-Rasmussen, T. Hedonism, Preferentialism, and Value Bearers. *The Journal of Value Inquiry* **2002**, *36*, 463–472.
34. Rutledge, R.B.; Skandali, N.; Dayan, P.; Dolan, R.J. A Computational and Neural Model of Momentary Subjective Well-being. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 12252–12257. doi:10.1073/pnas.1407535111.
35. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press, 1998.