Harsányi's Utilitarian Theorem

Kaarel Hänni

March 2023

1 Introduction

Harsányi's Utilitarian Theorem relates the decision-making of a collective of rational agents to the decision-makings of its constituent individuals. Harsányi originally framed the theorem in the context of welfare economics and social choice theory, in which case the theorem says that a rational group of agents' decision-making is given by maximizing an affine function of the welfares of the individuals. Two primary drivers of my interest in the theorem are links to agent foundations (hierarchical agency) and to ethics (moral uncertainty). This note is a self-contained[1] presentation of the theorem, including a succinct proof.

Before we get to this proof of Harsányi's Utilitarian Theorem, we first spend some time introducing terminology (Section 2) and proving the [Von Neumann]-Morgenstern Utility Theorem (Section 3) Readers familiar with the vNM theorem are encouraged to skip ahead to the discussion of Harsányi's Utilitarian Theorem in Section 4. Some philosophical context complements the math along the way.

2 Setup

Here, we set out the language of the vNM and Harsányi theorems. Our treatment applies to the case where there is a finite set of outcomes.[2] Let this set be $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$. Let us, say, think of each A_i as a complete deterministic world history, i.e. a 4D block of spacetime. Our agents have preferences between probability distributions on $\mathcal{A}$, i.e. preferences are defined for pairs of elements in $\Delta(\mathcal{A})$.[3][4] It will be convenient to represent the data of a probability distribution μ in vector form as

$$\boldsymbol{\mu} := (\mu(\{A_1\}), \mu(\{A_2\}), \dots, \mu(\{A_n\})).$$

[1] except for assuming some probability theory and linear algebra

[2] I believe an analogous story is true in more generality, but I have not investigated the more general variants in detail. For instance, I refer the interested reader to [3] for a treatment of the case with infinitely many outcomes.

[3] $\Delta(\mathcal{A})$ is the set of probability measures on the discrete σ-algebra $2^{\mathcal{A}}$, i.e. the set of all functions $\mu\colon 2^{\mathcal{A}} \to \mathbb{R}$ which satisfy $\mu(\emptyset) = 0$ and $\mu\left(2^{\mathcal{A}}\right) = 1$ and are additive under disjoint union.

[4] In this context, probability distributions are also often called *lotteries* or *prospects*.

Convex combinations of probability distributions are also probability distributions. For example, given $\mu, \nu \in \Delta(\mathcal{A})$, for any $p \in [0,1]$, we have that $\lambda = p\mu + (1-p)\nu$ is also a probability distribution. One can sample from λ by first flipping a biased coin with probability p of landing on heads. If it lands on heads, then sample from μ; if it lands on tails, then sample from ν.

Definition 2.1. A preference specification $\preceq$ over pairs of distributions is a subset of $\Delta(\mathcal{A}) \times \Delta(\mathcal{A})$, where we write $\mu \preceq \nu$ to mean that (μ, ν) is in the subset, and to intuitively mean that μ is no better than ν. We will take $\mu \succeq \nu$ to mean $\nu \preceq \mu$, we will take $\mu \sim \nu$ to mean that both $\mu \preceq \nu$ and $\nu \preceq \mu$, and we will take $\mu \prec \nu$ to mean that $\mu \preceq \nu$ but $\nu \not\preceq \mu$, and analogously define $\mu \succ \nu$.

Definition 2.2. We say a preference specification $\preceq$ is *vNM-rational* if it satisfies the following axioms:

Completeness: For any $\mu, \nu \in \Delta(A)$, we have that $\mu \preceq \nu$ or $\nu \preceq \mu$.

Transitivity: For any $\lambda, \mu, \nu \in \Delta(A)$, if $\lambda \preceq \mu$ and $\mu \preceq \nu$, then $\lambda \preceq \nu$.

Continuity: For any $\lambda, \mu, \nu \in \Delta(A)$, if $\lambda \preceq \mu \preceq \nu$, then there is a $p \in [0,1]$ such that $\mu \sim p\lambda + (1-p)\nu$

Independence: For any $\lambda, \mu, \nu \in \Delta(A)$ and $p \in (0,1]$, we have $\mu \preceq \nu$ if and only if $p\mu + (1-p)\lambda \preceq p\nu + (1-p)\lambda$.

Here are some immediate corollaries:

Transitivity′: If $\lambda \sim \mu$ and $\mu \sim \nu$, then $\lambda \sim \nu$.

Lifting: If $\mu \sim \mu'$ and $\nu \sim \nu'$, then $\mu \preceq \nu \iff \mu' \preceq \nu'$.

Independence′: For $p \in [0,1]$, if $\mu \sim \nu$, then $p\mu + (1-p)\lambda \sim p\nu + (1-p)\lambda$.

Interpolation: If $\mu \prec \nu$, then $p \leq q \iff (1-p)\mu + p\nu \preceq (1-q)\mu + q\nu$.

3 The vNM theorem

The vNM theorem tells us that a rational agent chooses as if it were maximizing the expectation of some quantity:

Theorem 1 ([Von Neumann]-Morgenstern Utility Theorem, 1947 [4])**.** *If an agent's preferences $\preceq$ are vNM-rational, then there is a function $u\colon \mathcal{A} \to \mathbb{R}$ which serves as its* utility function,[5] *by which we mean that*

$$\mu \preceq \nu \iff \mathbb{E}_\mu[u(A)] \leq \mathbb{E}_\nu[u(A)].$$

[5]Note that this is a purely behavioral condition – in particular, the theorem does not tell us that the agent has to internally be structured like a utility maximizer, or even that it has a concept of utility.

Proof. (From [5].) To construct such a function $u\colon \mathcal{A} \to \mathbb{R}$, we will first bring all distributions to a common scale. Let α_i be the probability distribution that assigns probability 1 to A_i, and assume WLOG that $\alpha_1 \preceq \ldots \preceq \alpha_n$. For all α_i, we have $\alpha_1 \preceq \alpha_i \preceq \alpha_n$, so by the Continuity Axiom, there is a $p_i \in [0,1]$ such that $\alpha_i \sim (1-p_i)\alpha_1 + p_i\alpha_n$.

Consider a distribution μ; we can write it as $\mu = \sum_{i=1}^n q_i\alpha_i$. Since $\alpha_i \cong (1-p_i)A_1 + p_iA_n$, Independence$'$ and Transitivity$'$ give that

$$\mu \sim \sum_{i=1}^n q_i\left((1-p_i)\alpha_1 + p_i\alpha_n\right) = \left(\sum_{i=1}^n q_i(1-p_i)\right)\alpha_1 + \left(\sum_{i=1}^n q_ip_i\right)\alpha_n =: \mu'.$$

We are now ready to construct the desired $u\colon \mathcal{A} \to \mathbb{R}$. If $\alpha_1 \sim \alpha_n$, then by Independence', $\mu' \sim \alpha_1$, in which case it follows from Transitivity$'$ that $\mu \sim \nu$ for any μ, ν, and thus the function $u = 0$ satisfies the property in the theorem statement. The remaining case is $\alpha_1 \prec \alpha_n$. In this case, define $u(A_i) = p_i$, and note that

$$\mu' = \left(1 - \mathbb{E}_\mu[u(A)]\right)\alpha_1 + \mathbb{E}_\mu[u(A)]\alpha_n.$$

Analogously, for another distribution ν, we have

$$\nu \sim \left(1 - \mathbb{E}_\nu[u(A)]\right)\alpha_1 + \mathbb{E}_\nu[u(A)]\alpha_n = \nu'.$$

By Interpolation, $\mathbb{E}_\mu[u(A)] \leq \mathbb{E}_\nu[u(A)] \iff \mu' \preceq \nu'$. And by Lifting, $\mu' \preceq \nu' \iff \mu \preceq \nu$, completing the proof. □

4 Harsányi's Utilitarian Theorem

Harsányi's Utilitarian Theorem connects the rational decision-making of a collective of rational agents to the utilities of the individual agents. Before stating the theorem, here is a list of cases the theorem can be applied to — that is, a list of cases where, plausibly, a collective of rational agents is making decisions on behalf of the individual rational agents:[6]

- different intelligent (alien) species agreeing to pursue common goals;
- the EU making decisions on behalf of its member states;
- a country making decisions on behalf of its citizens;
- a family making decisions on behalf of its members;

[6]Beware though that there are quite a few items on this list for which it is quite nontrivial how or whether the theorem applies. For example, in the moral uncertainty case, it is unclear if deontology corresponds to a set of rational preferences over ways the universe could be in some interesting sense. The same concern also applies to particular virtues in the virtue ethics example. Additionally, I think there are some wholes which might be "more than sums of their subagents". That is, I think it might make sense for some collectives to hold some preferences which are not held on behalf of their subagents. For example, it seems likely that this would be the case for decompositions of the human mind into subagents.

- a person making decisions on behalf of their (future) timeselves[7];
- a person making decisions combining verdicts of the different moral theories they are uncertain over;
- a person subscribing to virtue ethics making decisions that embody a set of virtues;
- an ethical theory grounding moral truth in concern for all rational agents, or for all sentient beings;
- an AI agent acting in the interest of a collection of humans.

Theorem 2 (Harsányi's Utilitarian Theorem, 1956 [2])**.** *Let $I = \{1, 2, \ldots, m\}$ index a set of agents, each with preferences; let $\succeq_i$ denote the preferences of $i \in I$. Suppose the set of agents taken as a unit also has preferences denoted $\succeq_I$. Suppose that the following three properties all hold:*

Individual Rationality: For each $i \in I$, $\preceq_i$ is vNM-rational.

Collective Rationality: Additionally, $\preceq_I$ is vNM-rational.

Pareto Optimality: [8][9] *If $\mu \sim_i \nu$ for every $i \in I$, then $\mu \sim_I \nu$.*

Then by the vNM theorem, there is a utility function $u_i \colon \mathcal{A} \to \mathbb{R}$ describing the preferences of each $i \in I$, as well as a utility function u_I describing the

[7] `https://www.lesswrong.com/posts/DdmScdukXBff5CbNS/a-gentle-primer-on-caring-including-in-strange-senses-with#Slicing_people_up_across_time`

[8] I will now explain why I call this condition Pareto Optimality. Note that it is implied by the property that if $\mu \succeq_i \nu$ for every $i \in I$, then $\mu \succeq_I \nu$. Given some additional assumptions (for instance that there is a pair of lotteries λ, κ such that $\lambda \succeq_i \kappa$ for all $i \in I$ with at least one strict inequality), this condition follows from the condition that if everyone weakly prefers μ to ν with at least one $i \in I$ strictly preferring μ to ν, then $\mu \succ_I \nu$. The reason to call this last condition Pareto Optimality is that if it were not the case that $\mu \succ_I \nu$, then the "collective policy" $\succ_I$ would not necessarily pick μ over ν when given the choice, and every $i \in I$ would weakly prefer, with one $i \in I$ strictly preferring, a collective policy which is the same except for definitely picking μ over ν. In other words, the collective policy would not be Pareto optimal. And conversely, if there is some other collective policy which every $i \in I$ would non-strictly prefer and some $i \in I$ would strictly prefer to $\succeq_I$ regardless of the choice between two lotteries faced, then there must be some μ, ν such that $\mu \succeq_i \nu$ for all $i \in I$, $\mu \succ_i \nu$ for some $i \in I$, and $\mu \preceq_I \nu$. (However, if one assumes some probability distribution on the set of choices between lotteries faced, then one can write down a more stringent Pareto condition. One can also think of this distinction as the given condition being Pareto Optimality for agents with Knightian uncertainty over the decision problem that is to be faced.) The initial condition that $[\forall i, \mu \sim_i \nu] \implies \mu \sim_I \nu$ is not equivalent to this last Pareto condition, for instance because it is also satisfied when $u_I = -\sum_{i \in I} u_i$ which definitely does not in general satisfy the Pareto condition, but since it is (given some reasonable additional assumptions) implied by a genuine Pareto condition and conceptually not that far from it, I think it still makes sense to call it Pareto Optimality.

[9] I think there are reasons to only accept this axiom when individual preferences are based in personal welfare only. See the counterexample provided in footnote 7 here: `https://www.lesswrong.com/posts/DdmScdukXBff5CbNS`.

preferences of the group. But what's more, there are coefficients $c_0, c_1, \ldots, c_k$ such that $u_I = c_0 + \sum_{i \in I} c_i u_i$.[10]

Proof. (Inspired by [1].) For each $i \in I$, we think of u_i as a vector, $\boldsymbol{u_i} = (u_i(A_1), u_i(A_2), \ldots, u_i(A_n))$, we let $\boldsymbol{u_0} = (1, 1, \ldots, 1) \in \mathbb{R}^n$, and consider the subspace $U = \text{span}\{\boldsymbol{u_0}, \boldsymbol{u_1}, \ldots, \boldsymbol{u_m}\}$. Define $\boldsymbol{u_I} = (u_I(A_1), u_I(A_2), \ldots, u_I(A_n))$, and consider $\boldsymbol{w} = \text{proj}_{U^\perp} \boldsymbol{u_I}$. In the rest of the argument, we will show that $\boldsymbol{w}$ is a "direction in the space of measures" which I "would care about" but which no $i \in I$ "would care about", implying that $\boldsymbol{w} = 0$ so as to not contradict Pareto Optimality.

Let μ be the uniform distribution on $\mathcal{A}$, i.e. $\boldsymbol{\mu} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, and consider the vector $\boldsymbol{\nu} = \boldsymbol{\mu} + \varepsilon \boldsymbol{w}$ where $\varepsilon > 0$ is chosen to be small enough to keep all entries positive. Since $\boldsymbol{w} \perp \boldsymbol{u_0}$, we have that $\boldsymbol{\nu}$ also corresponds to a probability distribution ν. For all $i \in I$, since $\boldsymbol{u_i} \perp \boldsymbol{w}$, we have $\mu \sim_i \nu$, and therefore using property 3,

$$\boldsymbol{u_I} \cdot \boldsymbol{\mu} = \mathbb{E}_\mu[u_I] = \mathbb{E}_\nu[u_I] = \boldsymbol{u_I} \cdot \boldsymbol{\nu} = \boldsymbol{u_I} \cdot \boldsymbol{\mu} + \varepsilon \boldsymbol{u_I} \cdot \boldsymbol{w},$$

whence $0 = \boldsymbol{u_I} \cdot \boldsymbol{w} = ||\text{proj}_{U^\perp} \boldsymbol{u_I}||^2$, implying $\text{proj}_{U^\perp} \boldsymbol{u_I} = 0$, so $\boldsymbol{u_I} \in U$. Therefore, $\boldsymbol{u_I} = c_0 \boldsymbol{u_0} + \sum_{i \in I} c_i \boldsymbol{u_i}$, and thus $u_I = c_0 + \sum_{i \in I} c_i u_i$.[11] □

Summary

In this paper, we discussed two canonical utility theorems of some relevance to ethics and to agent foundations. Firstly, we specified what it means for preferences to be rational. Secondly, we stated and proved the [Von Neumann]-Morgenstern Utility Theorem. Thirdly, we briefly discussed the possible philo-

[10] A fun aside: note that the special case with $I = \{1\}$ says that any two utility functions u, v representing preferences which are indifferent between the same pairs of lotteries are affine functions of each other, i.e. $v = a + bu$. In particular, any two utility functions representing the same preferences are thus affine functions of each other; in fact, it is easy to see that we can always choose $b > 0$ in this case, and that given $b > 0$, we indeed have that $a + bu$ represents the same preferences. In other words, the utility function for a given set of preferences is unique up to positive affine transformation.

[11] If we replace our Pareto Optimality condition with a slightly stronger one which says that if $\mu \preceq_i \nu$ for every i, then $\mu \preceq_I \nu$, then we can further take $c_i \geq 0$ for all $i \in I$. To see this, consider the convex cone $C = \left\{\sum_{i=0}^m c_i \boldsymbol{u_i} \mid c_0 \in \mathbb{R}, c_1, \ldots, c_m \in \mathbb{R}^{\geq 0}\right\}$. Say for a contradiction that $\boldsymbol{u_I} \notin C$. Then, since $C \subseteq \mathbb{R}^n$ is closed, there is a ball B of radius some $\varepsilon > 0$ around the point $\boldsymbol{u_I}$ which does not intersect C. Let $A \subseteq \mathbb{R}^n$ be the convex hull of $B \cup \{\boldsymbol{0}\}$. Note that since C is a cone, $B \cap C = \emptyset$ implies that $A \cap B = \{\boldsymbol{0}\}$. In particular, since $\boldsymbol{0}$ is not in the relative interior (`https://en.wikipedia.org/wiki/Relative_interior`) of A, it follows that A and B have disjoint relative interiors. By Separation theorem II from `https://en.wikipedia.org/wiki/Hyperplane_separation_theorem`, there is thus a nonzero vector $\mathbf{v}$ and a constant $k \in \mathbb{R}$ such that for all $\boldsymbol{a} \in A$, we have $\boldsymbol{a} \cdot \boldsymbol{v} \leq k$, and for all $\boldsymbol{c} \in C$, we have $\boldsymbol{c} \cdot \boldsymbol{v} \geq k$. From $0 \in A \cap C$, it follows that $c = 0$. From A containing a ball around $\boldsymbol{u_I}$, it follows that $\boldsymbol{u_I} \cdot \boldsymbol{v} < 0$. Note also that since $\pm \boldsymbol{u_0} \in C$, we must have $\pm \boldsymbol{u_0} \cdot \boldsymbol{v} \geq 0$, and thus $\boldsymbol{v}$ is orthogonal to $\boldsymbol{u_0}$, and thus $\boldsymbol{\nu} = \boldsymbol{\mu} + \varepsilon \boldsymbol{v}$ is a measure. In fact, $\boldsymbol{\nu}$ is weakly preferred over $\boldsymbol{\mu}$ by all $i \in I$ (by a calculation analogous to the one in the proof of the theorem), but $\boldsymbol{\nu} \prec_I \boldsymbol{\mu}$, a contradiction.

sophical significance of the Harsányi Utilitarian Theorem. Finally, we provided a semi-novel proof of the theorem.

Acknowledgments

Most of the thought that went into this piece happened at DASH — the Oxford Writing Workshop. I would like to thank its instructors and organizers, as well as its participants, for advice and insights. Some more optimization power went into this piece at the AI Safety Europe Retreat 2023, at which I gave a presentation on this material. I would like to thank the organizers and participants of AISER as well. And most importantly, I would like to thank Rio Popper for extensive feedback.

References

[1] Peter J. Hammond. "Harsanyi's Utilitarian Theorem: A Simpler Proof and Some Ethical Connotations". In: *Rational Interaction: Essays in Honor of John C. Harsanyi*. Ed. by Reinhard Selten. Remark: I actually used `https://web.stanford.edu/~hammond/HarsanyiFest.pdf`. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 305–319. ISBN: 978-3-662-09664-2. DOI: `10.1007/978-3-662-09664-2_17`. URL: `https://doi.org/10.1007/978-3-662-09664-2_17`.

[2] John C. Harsanyi. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility". In: *Journal of Political Economy* 63.4 (1955). Remark: I actually used `https://hceconomics.uchicago.edu/sites/default/files/pdf/events/Harsanyi_1955_JPE_v63_n4.pdf`, pp. 309–321. ISSN: 00223808, 1537534X. URL: `http://www.jstor.org/stable/1827128` (visited on 03/22/2023).

[3] I. N. Herstein and John Milnor. "An Axiomatic Approach to Measurable Utility". In: *Econometrica* 21.2 (1953), pp. 291–297. ISSN: 00129682, 14680262. URL: `http://www.jstor.org/stable/1905540` (visited on 04/30/2023).

[4] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Second edition. Remark: I did not actually look at this source, but I am pretty sure (because Wikipedia told me) that the theorem first appears in an appendix of this. Princeton University Press, 1947.

[5] Wikipedia contributors. *Von Neumann–Morgenstern utility theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 24-March-2023]. 2023. URL: `https://en.wikipedia.org/w/index.php?title=Von_Neumann%E2%80%93Morgenstern_utility_theorem&oldid=1136349097`.