Questions & suggestions welcome:
daniel.munoz@unc.edu

12 Open Questions about Multidimensional Value

Daniel Muñoz
UNC Chapel Hill
12/16/2024

If ethics has taught me anything, it's that values can be slippery. But one way to pin down a thing's value is to model it as a vector of real numbers, where each number measures how good or bad that thing is in a certain dimension (Muñoz, 2022, 2023, ms). If coffee is better in one way, tea in another, we might model their values as, say, (10, 5) and (5, 10). To compare values, we partially order the vector space—for example, with a lexical ordering, a dominance ordering, or simply by adding up the dimensions and comparing sums.

Such models, increasingly common in value theory, are part of a broader set of trends in ethics and economics. It's now common to use vectors to model welfare (Broome, 2004; Nebel, 2022), and to use similar techniques from social choice theory and decision theory when studying ethical value (Gustafsson, 2020; Hedden, forthcoming; Hedden & Muñoz, 2024; Henning, 2023; Lederman, 2023, forthcoming; Muñoz, 2024; Nebel, forthcoming; Steele, 2022).

Perhaps the most exciting thing about this recent wave of work is that it raises genuinely new questions—and lets us ask some old questions in a new way. Rare delights, in a field as old as ethics!

Here I want to lay out 12 of the most interesting open questions about modeling values as vectors. Five are "internal" questions about how to develop these models. Five are "external" questions about where the models might fall short. And two are about applications. (I also include "subquestions.")

Questions & suggestions welcome:
daniel.munoz@unc.edu

Internal questions

1. **How many dimensions of value are there?**

    a. Infinitely many? Countably many?

    b. Are the traditional values—knowledge, virtue, pleasure, etc.—single dimensions? Or umbrellas under which we will find many dimensions?

    c. Is there a finite set of finite principles partially ordering the vector space of values?[1]

2. **Can we know *a priori* what dimensions of value exist, or can we only know about some dimensions through experience?**

    a. For example, if there is a distinct dimension of value measuring the welfare of each *actual* individual, then we have to know who exists to know what value dimensions there are. This is an instance of a broader class of examples involving nonfungible particulars. Is this class indeed an example of *a posteriori* dimensions?

    b. Are there other examples?

[1] If there is *no* finite set of finite principles, then we are "particularists" about the ordering of vectors; see Dancy (2004).

3. **Can qualitative duplicates—things that are perfectly similar, but numerically distinct—differ in some dimension of value?**[2]

    a. We may also need to duplicate a thing's environment, if that can affect a thing's value. But suppose we do just that—then would the duplicate have equal value in all dimensions?

4. **What are the most plausible, general rules for modeling insensitivity to sweetening? (Chang, 2002; Hare, 2010)**

    a. What *kind* of rule makes ties sensitive to sweetening? The obvious example is a rule that tells us to add components and compare sums. But consider a standard "lexical" rule on which $(x_1, y_2) > (x_2, y_2)$ if and only if either $x_1 > x_2$, or $x_1 = x_2$ and $y_1 > y_2$. This rule, which does not involve any sort of adding, makes all ties sensitive to sweetening. But whereas an additive rule makes parity sensitive to sweetening, the lexical rule just bans parity: the only way to tie overall is to tie in all dimensions.
    b. Is there a useful typology of all rules that induce sensitivity?
    c. Analogous questions to 4a and 4b could be asked about other well-known tricky nontransitivities, such as the much-studied Repugnant Conclusion

[2] This question relates to the question about the value of particulars alluded to in 2a. On particulars as value bearers, see Cohen (2012) and Nebel (2021).

(Parfit, 1984). Is there a useful typology of all these cases?[3]

5. **Are value dimensions *anonymous*—in the sense that vectors would receive the same ranking even if we permuted the order of dimensions—or are some dimensions special?**[4]

    a. This raises a further question of whether there is a fact of the matter about which specific model is correct—or if instead the correct model is underdetermined by the facts. Consider two models assigning 2D vector values. On the first model, vectors are compared using ADDITION (an anonymous rule). On the second model, vectors are compared using a version of addition that gives the first dimension twice as much weight—but each thing we compare has only half the value along this dimension that it has according to the other model. (Value on the other dimension is the same in both models.) If it is indeterminate which model is correct, it is indeterminate whether the dimensions are anonymous.

---

[3] See, e.g., Bovens (2022). We may also generate typologies by looking at impossibility theorems which show that some principles rule out transitivity.

[4] Here I am drawing on May's (1952) notion of anonymity in social choice theory; see also Sen (2017: 5*). Lexical rules, which rank the dimensions in terms of priority, are a paradigm case of non-anonymity.

External Questions

6. **Can we think of absolute value terms, like "good," "bad," and "very good," as regions in a vector space?**

    a. Of course, which region "very good" refers to might depend on the context of conversation. But hold fixed the context. *Now* can we map terms to regions?

    b. Will there be plausible, nontrivial principles linking absolutes to comparatives? (For example, "If A is good, and B ties A, then B is not bad"?) How bad is it if the answer is "no"?[5]

7. **Why might some dimensions fail to have the structure of the real numbers?[6]**

    a. Are they bounded?

    b. Finite or only countably infinite?

    c. Can we not add them together? Not multiply by scalars?

---

[5] See Nebel (2018).

[6] This is related to, but not quite the same as, the traditional question of scale type. See Hedden and Nebel (forthcoming). (Even if values are measurable on a cardinal scale, that doesn't tell us whether they are, say, bounded from above.)

d. Might something fail to have any value, even a zero value, in some dimension? (Could a thing fail to have *any* value in *any* dimension?)

e. Do they not have the least upper bound property? (Note that *all* subsets of the reals have the least upper bound property, whereas plenty of subsets lack other properties characterize the reals—e.g. uncountability.)

8. Are there some dimensions that, unlike real numbers, themselves have multiple subdimensions?

a. Does *every* subdimension have subdimensions?

b. What is the structure of subdimensions? (Is the relation of being a subdimension transitive? Complete? If D and D^* are incomparable with respect to that relation, could they still share a subdimension?)

c. What combinations (or permutations) of subdimensions make for a superdimension of value?[7] Suppose d is a subdimension of D, and d^* of D^*,

[7] This is a case of Van Inwagen's (1990) "Special Composition Question" in mereology, the study of parts and wholes—from which we can borrow outlines of a possible debate. The universalist says that any combination of subdimensions forms a genuine dimension; the nihilist says that *no* combination forms a superdimension; and the commonsense view is that only some form a superdimension. The universalist seems to posit spurious superdimensions; the nihilist seems to miss out on genuine ones; and the commonsense view leaves us wondering, "Why just *these*?"

where *D* is not a subdimension of *D** or vice versa. Can there be a dimension *D*** that has *d* and *d** as subdimensions? Must there be?

d. Should we think of overall value as the ultimate superdimension?[8] Or should we think that "overall value" merely picks out different superdimensions depending on context?[9]

9. Does a thing have a fixed value along all dimensions, or does its value along some dimensions change depending on context?

a. Depending on context, a value might have *different dimensions*, or it might have *different values (in the same dimensions)*. This is the difference between saying "Equality is no longer a dimension of A's in this comparison" as opposed to "A receives a higher score on the dimension of equality in this comparison than it does in that one."

b. There may be several kinds of contextual influence. Does A's value depend on whether we are comparing it to B or to C? Does it change depending on which other options are on the menu?[10] Perhaps both: maybe, when D is on the menu, A's value is (1, 0) when compared to B and (0, 1) when

---

[8] Maybe value is "junky" (Schaffer 2010), and there is *no* ultimate superdimension.

[9] See Chang (2002, 2004) on "covering values."

[10] See the literature on Temkin's "essentially comparative view" (2012), which arguably includes both kinds of context-sensitivity.

compared to C, though with D off the menu, A's value is (2, 0) when compared to B and (0, 2) when compared to C.

c. Is there some contextual influence stemming from the "domain" of value at issue—aesthetic, moral, prudential, etc.?

10. Is overall value totally determined by value within dimensions?

a. This is the *dimensionalist* hypothesis.[11] One question is whether there might be multiple ways to make it precise. For example, there might be an absolute version ("How good a thing is overall supervenes on how good it is along every relevant dimension") and a comparative version ("How A compares to B overall supervenes on how A and B compare along every relevant dimension"). How we formulate dimensionalism may also depend on our conception of overall value, an issue raised by 8d.

b. Are there any counterexamples to dimensionalism?[12] Is it knowable *a priori* that we will never find counterexamples?

c. What familiar ethical theories entail dimensionalism? Do any theories rule it out? Are some theories compatible with dimensionalism and also with its negation?

[11] See Hedden & Nebel (forthcoming) and Hedden & Muñoz (2024).

[12] On potential counterexamples, see Hedden & Muñoz (2024) and consider 9c.

Applied Questions

11. What implications do vector models of value have for the philosophy of artificial intelligence?

a. It is possible that AIs deploy extremely high-dimensional representations of values that we cannot understand in any intuitive way.[13] Could we invent algorithms to map such representations to low-dimensional vector spaces with a more intuitive partial ordering? Could we define a useful error metric for such mappings? If so, we may be able to extract from AI models meaningful (though imperfect) justifications of their decisions.[14]
b. It is also possible, perhaps for reasons related to questions 6–9, that AIs do not represent values as fixed vectors of real numbers. In that case, given a certain evaluative question, might we still map whatever representation they do use to simple vector models and measure errors?

---

[13] It is controversial whether large language models (LLMs) like chatGPT 4.0 can represent *anything*—for example, whether the words they produce refer to things in the world (Lederman & Mahowald, 2024; Mandelkern & Linzen, 2024). But even if their words and internal states can refer, it is a further question *what* they refer to. On how to determine the beliefs of LLMs, see Herrmann and Levinstein (2024).

[14] On the right to such explanations, see Vredenburgh (2021).

12. Could vector modeling be a useful default approach to applied ethics in general?

a. Perhaps, if we are able to put vector values on things, there is no interesting work left to do. But might some important questions in applied ethics be naturally expressed with vectors? For example, perhaps the question of equality between species might be posed in terms of anonymous dimensions (see 5): does the dimension measuring goodness with respect to human welfare get extra weight?[15]
b. More pragmatically, what do we need to know *before* we can usefully bring in vectors? (Maybe it's never too early: failing to find the right vectors might help us find the gaps in our informal understanding of the moral issue at hand, just as a failed attempt to translate an argument into formal logic might reveal to us that we don't yet know how to formulate our premises.)

* * * * *

These questions are not completely open. But they are not completely closed, either, and I hope that, in 10 years, any advanced student of formal ethics will be able to understand them better than I understand them now.

[15] See Kagan (2019). This issue also raises issues of underdetermination (5a).

Works Cited

Bovens, Luc. (2022). Four Structures of Intransitive Preferences. In C.M. Melenovsky (Ed.), *Routledge Handbook of Philosophy, Politics and Economics* (81–93). New York: Routledge.

Broome, John. (2004). *Weighing Lives*. Oxford: Oxford University Press.

Chang, Ruth. (2002). The Possibility of Parity. *Ethics*, *112*(4), 659–688.

Chang, Ruth. (2004). 'All Things Considered'. *Philosophical Perspectives*, *18*(1), 1–22.

Cohen, G. A. (2012). Rescuing Conservatism: A Defense of Existing Value. In *Finding Oneself in the Other* (143–174). Princeton: Princeton University Press.

Dancy, Jonathan. (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

Gustafsson, Johan E. (2020). Population Axiology and the Possibility of a Fourth Category of Absolute Value. *Economics and Philosophy*, *36*(1), 81–110.

Hare, Caspar. (2010). Take the Sugar. *Analysis*, *70*(2), 237–247.

Hedden, Brian. (forthcoming). Parity and Pareto. *Philosophy and Phenomenological Research*.

Hedden, Brian, and Daniel Muñoz. (2024). Dimensions of Value. *Noûs*, *58*(2), 291–305.

Hedden, Brian, and Jacob M. Nebel. (forthcoming). Multidimensional Concepts and Disparate Scale Types. *Philosophical Review*.

Henning, Tim. (2023). Numbers without aggregation. *Noûs*. Early online.

Herrmann, Daniel A., and Benjamin A. Levinstein. (2024). Standards for Belief Representations in LLMs. *Minds and Machines*, *35*(1), 5.

Inwagen, Peter Van. (1990). *Material Beings*. Ithaca: Cornell University Press.

Lederman, Harvey. (2023). Incompleteness, Independence, and Negative Dominance. Retrieved from https://philarchive.org/rec/LEDIIA

Lederman, Harvey. (forthcoming). Of Marbles and Matchsticks. In Tamar Szabó Gendler, John Hawthorne, Julianne Chung & Alex Worsnip (Eds.), *Oxford Studies in Epistemology, Vol. 8*. Oxford: Oxford University Press.

Lederman, Harvey, and Mahowald, Kyle. (2023). Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs. *Transactions of the Association for Computational Linguistics*, 12, 1087–1103.

Mandelkern, Matthew, and Tal Linzen. (2024). *Do language models' words refere?* arXiv:2308.05576.

May, Kenneth O. (1952). A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision. *Econometrica*, *20*(4), 680–84.

Muñoz, Daniel. (2022). The Many, the Few, and the Nature of Value. *Ergo: An Open Access Journal of Philosophy*, *9*(4), 70–87.

Muñoz, Daniel. (2023). Sources of Transitivity. *Economics and Philosophy*, *39*(2), 285–306.

Muñoz, Daniel. (2024). Each counts for one. *Philosophical Studies*. Early online.

Muñoz, Daniel. (unpublished manuscript). Values as Vectors.

Nebel, Jacob M. (2018). The Good, the Bad, and the Transitivity of Better Than. *Noûs*, *52*(4), 874–899.

Nebel, Jacob M. (2021). Conservatisms About the Valuable. *Australasian Journal of Philosophy*, *100*(1), 180–194.

Nebel, Jacob M. (2022). Totalism Without Repugnance. In Jeff McMahan, Timothy Campbell, Ketan Ramakrishnan, & Jimmy Goodrich (Eds.), *Ethics and Existence: The Legacy of Derek Parfit* (200–231). Oxford: Oxford University Press.

Nebel, Jacob M. (forthcoming). Ethics Without Numbers. *Philosophy and Phenomenological Research*.

Parfit, Derek. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Schaffer, Jonathan. (2010). Monism: The Priority of the Whole. *Philosophical Review*, *119*(1), 31–76.

Sen, Amartya. (2017). *Collective Choice and Social Welfare: An Expanded Edition*. Cambridge: Harvard University Press.

Steele, Katie. (2022). Incommensurability that can(not) be ignored. In Henrik Andersson & Anders Herlitz (Eds.), *Value Incommensurability* (231–246). New York: Routledge.

Temkin, Larry S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Vredenburgh, Kate. (2021). The Right to Explanation. *Journal of Political Philosophy*, *30*(2), 209–229.