Agreement between moral models

Joe Roussos

#normativemodels #paperidea

Introduction

My topic here is the agreement or disagreement of theoretical frameworks in normative ethics, commonly called "moral theories". Classic examples include the disagreement between utilitarianism and Kantian ethics about what makes an action morally right, as well as their different judgements in stylised cases such as the trolley problem. The question I am interested in is what to make of disagreement or agreement between moral frameworks.

There are several nearby topics that this is relevant to. The first is moral disagreement, although that debate tends to focus on disagreement between *people*--either "the folk" or professional ethicists. While theoretical disagreements relate to disagreements between people (they may cause them or be the result of each disagreeing party's systematisation of their position), they are conceptually distinct.

The second relevant topic is moral uncertainty. That discussion concerns agents who are normatively uncertain, which is to say that even when all empirical facts are resolved they are uncertain about what they ought to do. In discussions of moral uncertainty, agents are often presented as being uncertain because they are uncertain about what the correct moral theory is, and because the moral theories on offer disagree. My setup is similar to that in much of the moral uncertainty debate, since I depart from a position in which there are several moral frameworks on the table, each of which we presumably have some interest or credence in, and which disagree.

What I will do is to apply a modelling lens to the question of theoretical disagreement. By a "modelling view" I mean the position that I recently defended (Roussos 2022): that it is fruitful for normative ethics to reconceptualise its theoretical frameworks as models rather than theories, where both "model" and "theory" are terms borrowed from science.

One way to think about this paper is the following: in his paper, Klemens says: nobody theorises convergence in ethics! But, he says in a footnote, it is all over philosophy of science. This talk is one window into that philosophy of science discussion, with some sketches of how to apply it to ethics.

1. What is a model? The pragmatic view

"Model" is a common and perhaps overused word. You probably think you know what it means. Nonetheless I will define it, since I use "model" in a specific sense. "Models are idealised representations, which form part of an indirect strategy of inquiry (called modelling)." (Roussos 2022,1)

- NB: This is *not* the sense of model familiar from logic: not the "model" in "model theory", "a theory is a collection of models", or "a model is a structure which realises a theory and makes all its sentences true". If you typically think of models as set-theoretic structures you will need to take this section as stipulating a new meaning for that term.

My favourite introductory example:
"What is a model? A favourite example will get us started, before the more detailed characterisation below. Imagine that I am studying the fish population in my local pond. I observe the fish feeding, breeding, and dying, for a few generations. I realise that the pond has a finite carrying capacity for fish, due to their needs for space and competition for food. I observe that the population this week depends positively on the population last week, but that as the population reaches the capacity of the pond, crowding hampers population growth. Reflecting on these patterns, I decide to use the following equation to predict changes in the fish population: $4N_{t+1} = N_t(1 - N_t)$, where N is the number of fish in the pond divided by the carrying capacity, and t is a time index counting months.
In so doing, I am modelling the fish population. This involves representing the fish population, in my case mathematically. Only certain features of the real pond and fish are represented, however; I have ignored the natural variation in fish size and reproduction. I have also ignored factors which I know to influence the population level of the actual pond, such as fishing. I treat time as discrete, and count in months. I make no claims that this equation describes fish growth everywhere: the form of the equation is chosen to fit the rate of reproduction of this population. The features that I will take as characteristic of modelling in this paper are these: (1) I *represent* the fish pond, in this case, using mathematics; (2) this representation is *idealised*: it leaves out some properties and adds in others which the real pond lacks; and (3) the idealised representation acts as a *proxy*, I study it to learn about the population in the pond." (Roussos 2022, 1)

What is part of the model?

- Background practice of mathematical modelling, along with tools for analysis---in this case the analysis of unstable different equations like the logistic model.

- Open question whether we should include in the description of the model: measurement practices and devices, approximation schemes, rules of thumb, theoretical assumptions about what is an interesting question in population biology, etc.
- The model is a kind of tool for doing inquiry

1.1. Representation

Scientific models are representational in two senses, often called representation-of and representation-as.

- Many scientific models are representations *of* real systems, which are called the "target" of the model. These can be either specific systems like my pond or kinds of system, like a predator-prey system. Without going too deeply into the theory of representation-of, we can note that it involves two systems of objects, one of which stands for or denotes the other. In the opening example, the mathematical variable N stood for the population density of the fish pond.
- Models also represent the world *as* being a certain way—typically a way which is simpler and different from how the world actually is. Some models don't have real systems as their targets, but they are nevertheless representations-as, just as a picture of a dragon is a kind of representation although there are no dragons.

1.2. Indirect

- Scientific inquiry with models is "indirect" in that the scientist spends their time working with and studying the model, as a proxy for the target system.
- Rather than counting fish in the pond, I manipulate the mathematical model and then make inferences about the fish pond.

1.3. Idealised

Models are characteristically *idealised*.

- Scientists typically cannot represent the systems they study completely accurately, either because the systems are too complex, or because their understanding is too limited, or because such a faithful representation would be intractable for analysis.
- So, in building their models, scientists leave out certain aspects of the system which they take to be irrelevant, and they represent the system as having properties that are different from its actual properties.
- These changes are called "idealisations" (Weisberg 2007a; Frigg and Hartmann 2018). Note that this term has no moral valence in science. Models are not thought to represent ideal systems in the sense of perfect or good systems.

Why idealise?

- Science makes frequent and seemingly ineliminable use of idealisation. But the presence of idealisations means that models contain known falsehoods.
- How do we square these facts?
- One recent thread emphasises that humans are inquirers with limited cognitive capacities, confronting a hugely complex reality. Idealisations enable us to manage that complexity, by isolating particular aspects of nature for study. When it works well, idealisation focuses attention on a real and important factor, sometimes by highlighting its salience to the researcher, sometimes by freeing it from its interactions with other factors. Clearly, not all idealisation is good and recent work has focused on characterising when it works well.
- For our purposes, the important lesson from that literature is that the success conditions are relative to the **purposes** of the inquiry, the **inquirer's capabilities**, and **the system being studied**.

Implication of this view: purpose-specificity and truth-aptness

So, models have these distortions which are pragmatically justified relative to a purpose. A user, a context, and a purpose make certain idealisations justified.

- First, models are not themselves candidates for truth. They are tools. They contain falsehoods. They aren't intended to be 1-1 representations
- Second, philosophers have argued that the success condition for scientific models is not truth per se, but adequacy for purpose (notably Parker 2009). This does not remove truth from the picture: it remains the ultimate goal of inquiry. But models have different immediate aims, against which they are evaluated. Truth is approached more indirectly, through a series of model-based inquiries which delivery important insights about limited domains, or particular questions.

2. Models in normative ethics

I previously proposed that it is fruitful to view theoretical frameworks in ethics as models. I don't have time to go through that argument fully, so here is a brief summary.

2.1. Broad picture

- **Observations**: observations of moral life, and our moral judgements. Like many natural domains, the ethical domain is extremely complex and we have only partial information about it.
- **Data**: Stable considered judgements, intuitions. "Cleaned-up" observations.

- **Initial analysis**: Ethicists discern certain patterns amongst these, which they investigate, seeking eventually to systematise them. There may be some empirical regularities (e.g., common judgements, apparent norms), which we aim to explain by the introduction of theoretical concepts (e.g., precisified notions of duty, or welfare). But the domain is complex, patterns are hard to discern, and the data often seem contradictory, and so it is difficult to "read off" moral laws from the data.
- Different goals:
  - Theory: set of principles which are universal and true
  - Model: idealised representation used in indirect inquiry. Can stand in several relations to theory

Idealisations and purposes

- "Providing a decision procedure" and "providing a criterion of rightness" are two purposes which are classically distinguished in ethics. To this we might add the much simpler "rendering a judgement".
- There are also domains of ethics with prima facie different constraints, goals, and relevant factors:
  - Relations between intimates
  - Distribution of scarce resources
    - State level
    - Hospital triage
    - Stranded on a desert island
  - Decision-making on behalf of others
    - Limited autonomy situations
    - State level

2.2. Models as mediators

- **Stepping stones**: Build a model to help test out different principles, in the process of developing a theory
- **Bridges**: Models connect an existing high-level theory, like utilitarianism, with a particular domain.
  - In science: done by drawing on elements which are not part of the theory, including empirical information about the domain, approximation techniques, and diagrammatic methods.
  - In ethics: similar!
    - Question, or domain of interest, sets the scope of the inquiry. Examples of such domains are distributive questions for social planners, or duties of care. Work done in one of these domains will not usually be expected to

apply to the other, even if the philosopher doing the work thinks that a Kantian analysis is best in both cases.
    - This kind of work also involves idealisations, of both the leaving out and distorting kinds.
    - The ethicist studies a situation, or group of people, or situation, which is different in important ways from any real situation. Work of this kind therefore has the indirect nature which is characteristic of modelling.
    - The work might incorporate constraints drawn from real-world considerations and tools ranging from familiar test cases to diagrams and tables. All of these help to connect the content of the theory (a set of principles) to the domain being studied.

2.3. Models without theory

- For ethicists who are skeptical of particular theories, or the project of moral theory as a whole, ethical models might play this role.
- Another case of relatively theory-free modelling is in mid-level domains which are hard to connect to fundamental theories. Philosophers working on mid-level questions might find it simpler to work directly in the language of their level, rather than seeking connections with the language of the available theories. e.g., the "polluter pays principle"

2.4. "Theories" as models

- We have no theory, and what we call a theory (e.g., utilitarianism) is better understood as a model.
- e.g., Utilitarianism
    - Intended domain: Large-scale, interpersonal, distributive questions
    - Intended use: not a decision procedure
    - Idealisations: neglects considerations of partiality, other nonconsequentialist considerations
- e.g., Scanlon's contractualism. "Justifiability to each"
    - Intended domain: small-scale interpersonal questions (why: we consider others moral positions)
    - Idealisations: focuses only on individual reasons to reject, ignores "the goodness of outcomes" as a moral factor in itself
    - Representation clearly different, focuses on claims

Modelling view proposes that the criteria of rightness that you get from a moral model might be context and purpose specific, in the fuzzy way that models are.

On this view one would not have (unqualified) credence in a "theory" like utilitarianism, because it is really a model. This blocks various inferences.

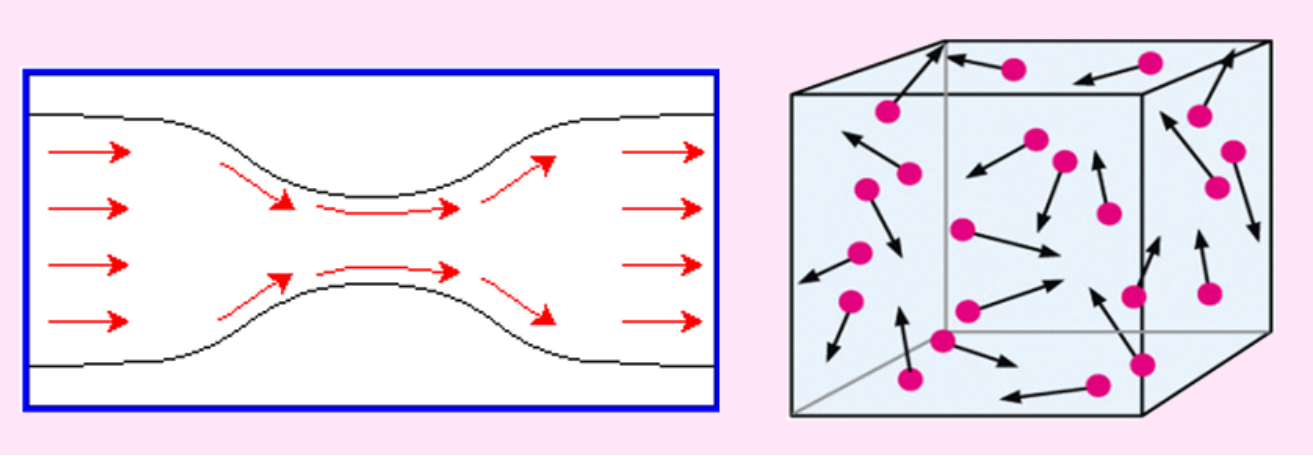
- e.g., I have credence p in theory T and theory T delivers judgement J, so I have credence p in J

3. Disagreement

So, suppose that the main frameworks available in ethics are models rather than theories. What does this mean?

3.1. Purpose-specificity and disagreement

- Models are purpose-specific tools of inquiry. The purposes of inquiry, inquirer's capabilities, and eventual idealisations together set a domain of application for the model—outside of which it should not be expected to work well (Teller 2001; Weisberg 2007b).
- This feature of modelling explains why we encounter multiple, disagreeing scientific models of the same phenomenon.
- Teller illustrates this with an example of two models of water.
    - The first is interested in the flow of water and wave propagation, and it represents the liquid as a continuous incompressible medium.
    - The second is interested in explaining diffusion, say of a drop of ink in water. It represents water as a collection of discrete particles in thermal motion.
    - Each is similar to water in the respects that are relevant to its purpose, but the two models look very different (Teller 2001, 401). Each is highly successful at its purpose, i.e., prediction of the relevant kind of behaviour, and their respective idealisations work well within their domain.
    - But clearly they contradict one another: one says that water has particles, the other says it does not. The lesson is that neither should be thought to provide a definite characterisation of water, and our understanding of water is enhanced by having both available.



ASIDE: Counterexamples

- Counterexamples count against a model directly only when they are within its scope. It it therefore of first importance to delineate these scopes, explicitly and upfront, when we model.
- Counterexamples from outside a model's scope can count against it indirectly, when models are being compared. If a second model performs better on the shared purpose and has wider scope, then it will be favoured.

"Disagreement" is a complex business.

- It isn't disagreement unless one fixes all relevant factors: purposes, domains, capabilities
- Once relevant factors are fixed, what can we say about disagreement?

3.2. Disagreement and uncertainty

One common "use" of model disagreement is as a representation of uncertainty. The logic is this:

- We have a group of models, $m_1, \ldots, m_n$ developed for roughly similar purposes
- The differences between them represent different choices at relatively unconstrained choice-points in model construction. i.e., The modellers used different theories, made different assumptions, used different idealisations, etc.
- The set of available options at each of these points is due to our uncertainty (about the right theory, the best technique, the appropriate assumption, etc.)
- Thus, the diversity of our model results represents (partially and incompletely) our uncertainty
- One popular concept, for situations of high uncertainty (when it is difficult to know how to weigh or compare models): the range of model results is a "non-discountable region"--we can't exclude that the right answer is within this range. That's it.

Note the difference in framing to the standard framing of moral uncertainty.

- In MU, utilitarianism etc are theories and are assigned credences. Approaches like "maximise expected choiceworthiness" use these credences as weights in the evaluation of actions
- Models are not candidates for truth and so a model cannot be an object of credence *tout court*. It doesn't matter if you're a hardnosed moral realist, these just aren't truth-apt. They're more like hammers than sentences.
- Model results can, but these credences will vary from result to result, depending on features like the context, how it matches the model's intended purpose, beliefs about the role idealisations are playing, etc.
- Model results are claims, and different models can output the same result. If the credences are in claims, the structure of the moral uncertainty problem looks a bit different
    - Not all aspects of model output are meant to be read literally. There are artefacts, there is excess precision, etc. Comparing the outputs of "theories" directly--some give cardinal scale numbers, etc.--might be misleading

4. Agreement between models in science

4.1. Spurious agreement

Agreement outside of a model's domain

- Models come with intended purposes and domains---their "home turf"
- Say we have three models m_1, m_2, m_3 with domains D_1, D_2, D_3, which overlap. If the agreed-upon result R is in their intersection $D_1 \cap D_2 \cap D_3$ then there may be reason to increase confidence in R. If the result is in D_1 but not the others, then it is unclear whether models 2 and 3 add anything. At minimum, some work needs to be done to describe how these models function outside of their home domains, and to investigate whether there are distortions/sources of error.
    - If the diffusion model of water agrees with the flow model of water about some flow properties, this might not be any reason for increased confidence in that property.

Agreement on the data

- Regurgitation of the data: if a model produces some claim, which is true, this seems to confirm the model. However, if that claim *just is* a piece of data used to construct the model, this confirmatory effect vanishes. Sometimes we distinguish between

    - Verification of a model: checking that you didn't make mistakes by, inter alia, getting it to reproduce data used in its construction
    - Validation of a model: checking whether the model is a good representation of the system for its purpose by, inter alia, predicting unobserved behaviour of the target and then measuring the target to confirm the prediction
- What can we make of this distinction in ethics?
    - Which intuitions are used to construct ethical models and which are "unobserved" and can be used as data for "predictive testing"?
    - There is a particular problem affecting our ability to assess this for moral models. Moral models are prime examples of what philosophers of science have described as "reflexive prediction" and "performative models"
        - Models don't just describe/predict, they also influence. Classic examples are economic models, which change how individuals, firms and states behave and thus alter the markets that the model set out to describe.
        - Moral behaviour has changed over time, partly in conversation with ethical theory
    - One might attempt to construct a minimal version of, e.g., utilitarianism using only a few intuitions. It might be difficult to ensure that this is really happening unless one produces a formal proof where the data can be formalised into conditions, and the moral principles derived from them logically. Something like Harsanyi's theorem.

4.2. Robustness

Primary question in philosophy of scientific models: is agreement between models confirmatory? i.e., When models $1, \ldots, n$ agree on a result R, is R more likely to be true than it would have been if only model 1 gave that result?

- Many people think so (scientists, philosophers). Such results are called "robust". (e.g. Weisberg 2006, 2013;1 Lloyd 2009, 2015; Kuorikoski et al. 2010, 2012)
- But, it is very hard to explain why. The literature is largely a story of failed attempts (Kuorikoski et al. vs. Harris, Levins vs. Orzack & Sober, Harris & Frigg, etc.)

The core idea is often traced to this quotation from Levins:

> "Even the most flexible models have artificial assumptions. There is always room for doubt as to whether a result depends on the essentials of a model or on the details of the simplifying assumptions. [. . .] Therefore, we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results

```
we have what we can call a robust theorem which is relatively free of the
details of the model. Hence our truth is the intersection of independent
lies." (Levins 1966, p. 423)
```

The basic robustness idea is that it seems *miraculous* for the same result to appear multiple times if it is false. Inference to the best explanation: the robust result is true! There are two ways that people propose to use robustness.

1. To discover "robust theorems" or properties. Roughly, we observe agreement on some property, factor, relationship and then infer from this a claim about the target system.
2. To confirm hypotheses about the target system. A single model producing a result which is consistent with a hypothesis corroborates it to some degree. Many models doing so appears to boost that confirmation.

There is a basic problem with this, however.

1. Models contain falsehoods (idealisations, simplifications) and unjustified assumptions
2. Each model's result is consequence of all of its parts
3. So each result is a consequence of some falsehoods

What is needed is an explanation of how model agreement overcomes this problem. There are many ways that people have tried to make this case, but they all seem to rely on assumptions about probabilistic independence. Here are two examples

1. Measurement, methods and statistical sampling
2. Wisdom of crowds

4.2.0 Inferential robustness

There are many attempts to formulate logical or probabilistic arguments which secure robustness.
e.g., Orzack and Sober

1. $m_1 \vee m_2 \vee \ldots \vee m_n$
2. $m_i \rightarrow R$ for each i
3. So, R
   What does (1) mean? It can't mean this model is true, since the model isn't a sentence, set of sentences, etc. Perhaps it is a statement about the model producing the correct result in this context. Since each model can derive the result, (2) asserts that if m_i then the result is true for each model. The problem is: this

requires us to be certain that one of our theories is the correct one. How plausible is that?

4.2.1 Measurement and sampling

There is a well-worn idea that methodological triangulation is epistemically valuable. If one arrives at the same result from two different methods of investigation, this renders the result less likely to be a mistake.

- Independent errors are at the core of methodological triangulation. One way of securing this is to exploit different causal mechanisms.

Basic statistics:

- **Population**: a large collection of objects
- **Sample**: a subset of the population
  - Samples are typically selected according to some properties of the members of the population
  - **Representative sample**: one chosen using a selection process that does not depend other properties of the population. - For example, a representative sample of English voters in the 2017 election might consist of a randomly sampled set of 10,000 of the English people who voted in the 2017 election. On the other hand, a sample chosen from English Twitter users who voted in the 2017 election may not be unbiased, since many English voters are not on Twitter.
  - **Random sample:** one in which every member of the population has a non-zero probability of being selected, according to a known (or determinable) distribution
- **Estimator:** a statistic (a function of the sample data) that produces an estimate of a desired quantity
  - Random samples make it possible to produce unbiased estimates of population properties, as we can weigh the properties of elements of the sample according to their probability of selection to be in the sample.
  - The sample mean of a random sample is an estimator of the mean of the population. It is an unbiased estimator, as the expected value of the sample mean is the population mean.

Two routes to analogising between models and samples

1. Model results are like measurements and thus the collection of those results is like a sample.
2. Model building is like sampling the population of models

Problem: Collections of models are **not** samples

- What would the population be? Do we want to know its mean?
  - Population = Models which actually exist: not helpful
  - Population = Possible models. Why believe that its mean is the truth?
- Actual procedure isn't sampling, it is construction of a few things which occurred to us
  "There is no reason to believe model generation—a process carried out by scientists who know one another and work in a particular disciplinary matrix—will meet the technical definition for a random sample: a random sample is one in which every element of the population has a non-zero probability of being selected as a member of the sample, according to a probability measure on the population that is either known or can be determined. It is implausible that the relevant scientists are equally likely to generate each of all the plausible models, or that we could construct a distribution describing their probabilities of “selecting” particular models."

4.2.2. Wisdom of crowds

Recall Condorcet's Jury Theorem

- Suppose we are deciding on the truth of a proposition and wish to use the votes of a set of people to do so
- Assume (1) their votes are independent, and (2) each is >50% likely to get the truth-value correct
- The Theorem says that the larger the number of votes, the higher the probability that the majority is correct, and as the number of voters increases that probability tends to 1.

Two options for applying this to models in order to make model agreement analogous to majority voting in the CJT:

1. Model results ≈ votes, average ≈ majority position
2. Turn each model result into a categorical statement (e.g., true value lies above some threshold), then determine the literal majority position

Problem: Do models meet the conditions of the theorem? Independence and competence?

- It is hard to identify whether people or models give probabilistically independent answers. Note that they can't be totally independent since they aren't randomising devices, they should be sensitive to what the right answer is. But conditional on

the right answer, they should be independent which roughly means that their errors should be independent

- Science: Modellers share knowledge, disciplinary backgrounds, etc. Not plausible their models are independent. Often unclear whether individual models are Competent and robustness analysis cannot safely presume this
- Ethics: What would it mean for ethical models to be independent?

5. Conclusion

The purpose of this talk is to provide you with food for thought, by noting that it is fruitful to view ethics as in the business of modelling and that there is a large literature on agreement and disagreement between models in science.