THE PARLIAMENTARY MODEL OF MORAL UNCERTAINTY

HILARY GREAVES, OXFORD

EMPIRICAL UNCERTAINTY AND MORAL UNCERTAINTY

- Empirical uncertainty, prudential case: You are uncertain whether or not it will rain. Should you take an umbrella?
  - Standard answer: Maximise expected utility
- Moral uncertainty: You are uncertain whether eating meat is permissible. Should you eat meat?
  - Near-standard answer: Maximise *expected choiceworthiness* (MEC)

OUTLINE

1. ‘Maximise expected choiceworthiness’ (MEC), objections, and extant alternatives
2. The parliamentary model: Clarifying the model, and some examples
3. Structural features of the parliamentary model, and comparison to MEC
4. Conclusions

1.1 OBJECTIONS TO MEC

- The problem of intertheoretic comparisons (e.g. Hudson1989, Gracely1996, Broome 2012)
- Inability to accommodate moral theories that have 'awkward structures'
  - E.g. widespread incomparability, cyclic choiceworthiness relation
- Inability to accommodate 'hedging'
  - Sometimes it seems(?) appropriate to perform an 'intermediate action', even in a constant-marginal-returns setting. But MEC cannot recommend this. (TBC)

1.2 SOME ALTERNATIVES TO MEC

- Monistic externalism: One ought to do whatever is required by the moral theory that is *in fact correct*, end of story. (Harman 2011, Weatherson 2014, Mason 2015)
  - Insufficiently action-guiding (like monistic objectivism in the empirical case).
- Alternative internalist accounts, e.g. 'my favourite theory': One ought to do whatever is required by the moral theory one has highest credence in. (Gracely 1996, Lockhart 2000, Gustafsson and Torpman 2014)
  - (Among other objections,) cannot capture the importance of *relative stakes*…

1.3 THE INTUITION TOWARDS RESPECTING RELATIVE STAKES

- Empirical case: You're uncertain whether any potential bike thieves will pass this way in the next hour. Should you lock up your bike?
- Moral case: You're 90% sure that eating meat is permissible, but you have 10% credence that it's seriously morally wrong. Should you eat meat?
- The intuition towards respecting relative stakes, roughly: It's often appropriate to do something that is very likely to be worse, if (if it's worse it's *only a bit* worse, and if it's better it's *much* better).
- We/I seek: an internalist theory that respects relative stakes, but that is otherwise better than/different from MEC.

2.1 THE PARLIAMENTARY MODEL

- A suggestion (Bostrom 2009): "Suppose that you have a set of mutually exclusive moral theories, and that you assign each of these some probability. Now imagine that each of these theories gets to send some number of delegates to The Parliament. The number of delegates each theory gets to send is proportional to the probability of the theory. Then the delegates bargain with one another for support on various issues; and the Parliament reaches a decision by the delegates voting. What you should do is act according to the decisions of this imaginary Parliament."
- As it stands, the 'parliamentary model' is underspecified: What is the bargaining procedure/the voting procedure?

2.2 CLARIFYING THE MODEL: THE NASH BARGAINING SOLUTION

- Assume the (asymmetric) *Nash Bargaining Solution* (NBS), i.e. choose x to maximise

$$\prod_i (u_i(x) - u_i(d))^{a_i},$$

where: a_i is the agent's credence in theory i, d is the 'disagreement point', and x ranges over (pure and mixed) acts that are weakly Pareto superior to d.

- Justification: The NBS (i) is the unique subgame-perfect equilibrium of various sensible-looking bargaining procedures (Binmore, Rubinstein and Wolinsky 1986; Suh and Wen 2003), and (ii) is characterised by various nice-looking sets of axioms (Nash 1953; Lensberg 1988).
- NB The selected act x is independent of how the utility functions for each theory are (separately) scaled.
- Disagreement point, for the moral-uncertainty case: The mixed act in which the probability of any given pure act is equal to the agent's credence that that (pure) act is *best*.

2.3 MORAL UNCERTAINTY (I): VEGETARIANISM

- Suppose you have credence 90% that eating meat is morally permissible and marginally preferable to refraining from eating meat, but 10% credence that eating meat is seriously wrong. What should you do?
- MEC: (Probably) refrain.
- NBS: Perform a mixed act corresponding to eating meat with 90% probability. (There's no act that is Pareto superior to the disagreement point.)
  - The NBS approach always recommends the disagreement point *in scenarios that only contain two pure acts*.

2.4 MORAL UNCERTAINTY (II): UTILITARIANISM AND COMMON-SENSE MORALITY

- Suppose you have credence 50% in each of two theories: (1) a utilitarian theory, and (2) a 'common-sense' theory according to which (i) killing is significantly worse than letting die, and (ii) it's marginally better to support the local needy than to support x-risk mitigation.
- Suppose you face *two simultaneous* choices: whether to kill 1 or instead let 2 die, and whether to donate to the local needy or instead to x-risk mitigation.
- Suppose that the moral values (utilitarian, common-sense) are given by

	Kill 1	**Let 2 die**
Donate to x-risk	(+10, -10)	(+9, +9)
Support local needy	(-9, -9)	(-10, +10)

NBS

3.1 STRUCTURAL FEATURES OF THE NBS APPROACH

- Use of intertheoretic comparisons
- Respecting relative stakes
- Distinctness from MEC
- Ability to incorporate 'structurally awkward' theories
- Hedging

3.2 RESPECTING RELATIVE STAKES

- What we wanted: Under moral uncertainty, we often defer to the theory according to which the choice at hand is *high-stakes*, even if that is a theory we have somewhat lower credence in.
- The NBS approach gives us this *when there are many issues on the table* (and thus many available alternatives). But not when there are only two available (pure) acts.
- How this comes about (heuristically):
  - *Of course* the NBS approach can't accommodate (type I) hedging in two-pure-act cases – no intertheoretic comparisons, hence no recognition of different relative stakes across theories.
  - In many-alternative cases, the NBS approach takes account of how high the stakes are on one 'issue' *relative to* how high the same theory takes the stakes to be on other 'issues'. (Like 'structural' approaches to intertheoretic comparisons within MEC.)

3.3 DISTINCTNESS FROM MEC

- Suspicion: Is the NBS approach equivalent to a particular version of MEC, with some 'structural' way of setting the intertheoretic comparisons?
  - E.g. Variance normalisation: equalise the variance of possible moral values, across all moral theories. (Cotton-Barratt, Ord and MacAskill (MS))
- Reply: No. The NBS approach violates Independence/Sure-Thing: e.g. it will sometimes strictly prefer a mixed act over either/any of the pure acts involved.
  - Q: Is this an objection to the NBS approach? (Not obvious to me.)

3.4 (IN)ABILITY TO INCORPORATE 'STRUCTURALLY AWKWARD' THEORIES

- Some theories postulate a structure that is inconsistent with maximising EU: e.g., theories with cyclic choiceworthiness orderings.
  - (Less problematic: non-consequentialist theories; purely ordinal theories.)
- MEC cannot accommodate any such theories.
- But neither can NBS (for basically the same reasons).

3.5 HEDGING

- Fix two pure acts A, B. Suppose that for every $\lambda \in (0,1)$ there exists an 'intermediate' pure act $\lambda(A,B) \equiv \lambda A + (1-\lambda)B$, such that for each theory i, $u_i(\lambda(A,B)) = \lambda u_i(A) + (1-\lambda) u_i(B)$.
- Note that in this situation, an MEC approach can only recommend extremal acts (except in the degenerate case in which A, B have equal expected value).
- The NBS might strictly prefer some 'intermediate' act.
- This will initially look like an advantage of NBS over MEC (but ultimately I'll argue it isn't).

3.6 EXAMPLE: SPLITTING ONE'S PHILANTHROPIC BUDGET

- Suppose one has a fixed philanthropic budget, and seeks to maximise the amount of good done, but because of fundamental normative uncertainty, is uncertain which of two potential recipients is the more cost-effective (e.g. the top global poverty charity or the top animal welfare charity).
- A very natural response to this situation is to *split* one's donations.
- As noted, MEC apparently cannot recommend this 'intermediate act'. (Snowden (MS))
- The NBS does correspond to a point at which the agent splits her philanthropic budget, in proportion to her credences in the respective theories.

3.7 Q: DOES THE ABILITY TO RECOMMEND HEDGING COUNT IN FAVOUR OF NBS OVER MEC?

- I don't think so:
  - The intuitions in support of hedging apply just as much in the case of empirical uncertainty as in the case of moral uncertainty, but NBS can recommend hedging only in the latter case.
  - There *are* ways for MEC to recommend hedging – via concave utility functions (diminishing marginal returns).
  - In the linear model outlined, the NBS approach is indifferent between (i) hedging and (ii) choosing a *mixed act* with probabilities equal to one's credences in the relevant theories. But the pro-hedging intuition does not support mixed-outcome 'hedging'.

CONCLUSIONS

- So far, the NBS approach looks if anything slightly inferior to MEC:
  - Unlike MEC, the NBS approach makes no use of intertheoretic comparisons. However, it faces many of the same issues as structural approaches to intertheoretic comparisons within the MEC approach (in particular, sensitivity to what we take the full space of options to be).
  - NBS fares no better than MEC in dealing with theories that are incompatible with MEC's structure.
  - NBS recommends hedging in situations in which MEC does not and in which *superficially* we might want such hedging, but overall the NBS approach actually seems inferior to MEC on this score.
  - (And NBS violates Independence.)
- However: I've only investigated the NBS. But the NBS is not all of bargaining theory. It would be interesting to see whether any alternative bargaining/voting model performs relevantly differently.

REFERENCES

- Binmore, K., A. Rubinstein and Wolinsky (1986). "The Nash bargaining solution in economic modelling."
- Bostrom, N. (2009, blog post) "Moral Uncertainty – towards a solution?"
- Broome, J. (2012) "Climate matters."
- Cotton-Barratt, O., W. MacAskill and T. Ord. (MS) "Normative uncertainty, intertheoretic comparisons, and variance normalisation."
- Gracely, E. (1996) "On the noncomparability of judgments made by different ethical theories."
- Gustafsson, J. and O. Torpman. (2014) "In defence of My Favourite Theory."
- Harman, E. (2011) "Does moral ignorance excuplate?"
- Hudson, J. (1989) "Subjectivization in ethics."
- Lensberg, T. (1988). "Stability and the Nash solution."
- Lockhart, T. (2000) *Moral uncertainty and its consequences*.
- Mason, E. (2015) "Moral ignorance and blameworthiness."
- Nash, J. (1953) "Two person cooperative games."
- Snowden, J. (MS) "Donating to multiple charities."
- Suh, S.-C. and Q. Wen. (2003) "Multi-agent bilateral bargaining and the Nash bargaining solution."
- Weatherson, B. (2014) "Running risks morally."

OPEN PROBLEMS/QUESTIONS

- The NBS approach falls near-silent as soon as the agent has non-zero credence in a completely uniform theory, i.e. a theory that assigns equal value to all possible outcomes. This is bad.
  - Suggests a pertinent difference between the bargaining interpretation vs. the moral-uncertainty interpretation of 'bargaining theory'.