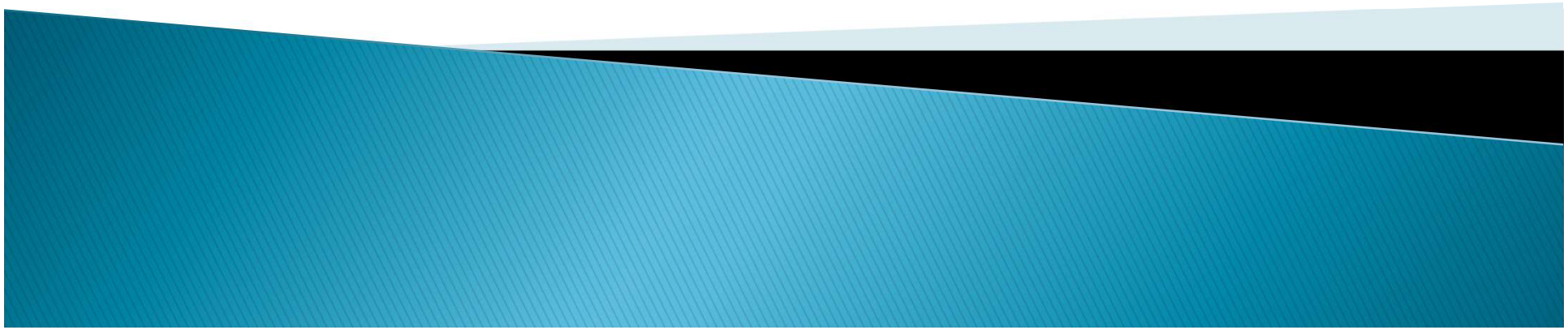
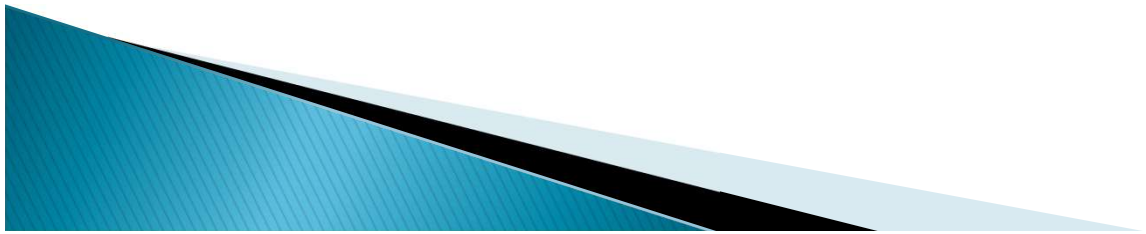
Cluelessness

Hilary Greaves (Oxford)



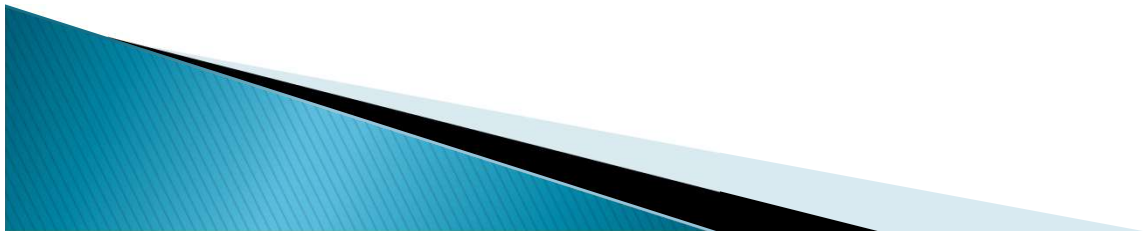
Cluelessness: The basic idea

- According to any plausible theory of morality (or rationality), consequences are important.
  - For decision-making, and for evaluation.
- But perhaps we can never know, *or even 'have the faintest idea'*, which of two actions will in the end have the better overall consequences (or which is *likely* to, etc.).
- If so, then… [some disaster for moral decision-making, moral evaluation, practical rationality].
  - Certainly given a consequentialist moral theory. But the problem is just as bad for other (plausible) theories too.



Outline

1. Cluelessness about objective betterness
2. Cluelessness about subjective betterness
3. Lenman's objection: The Principle of Indifference
4. The 'new problem of cluelessness'
5. The nature of cluelessness
6. Mundane cluelessness
7. Conclusions

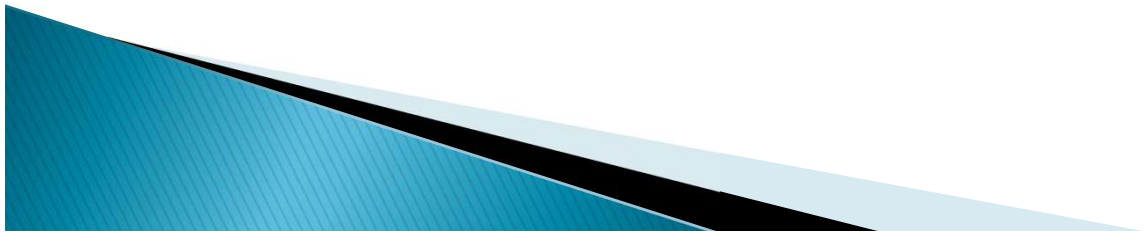


1. Cluelessness about objective betterness

- **(OB: Criterion of objective c-betterness)** A_1 is *objectively c-better* than A_2 iff the consequences of A_1 are better than those of A_2.
- **(CW_o: Cluelessness Worry$_o$)** We can never have 'even the faintest idea', for any given pair of acts (A_1, A_2), whether or not A_1 is objectively c-better than A_2.
  - Because of unforeseeable (indirect, longer-term) effects.
- **(NR_o (Non-reversal$_o$))** The net effect of taking into account the unforeseeable effects *would not reverse* the objective c-betterness judgments that we reach based on the foreseeable effects alone.
  - *If* we are justified in assuming (NR_o), this would rebut the cluelessness worry (CW_o). But are we??

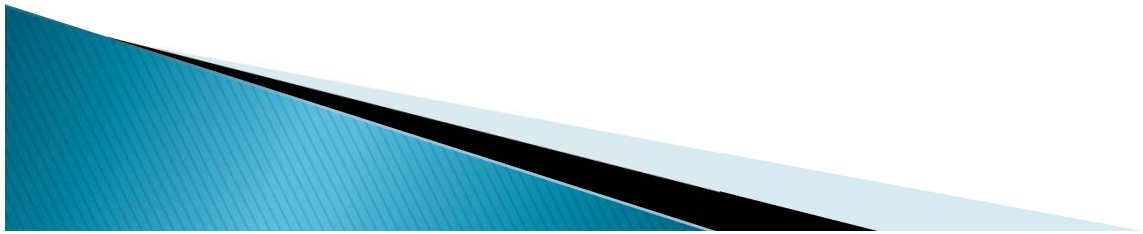
Avoiding cluelessness about objective betterness (I): Ripples on a pond

- The 'ripples on the pond postulate' (Moore, Smart): effects remote in time/space for the action become increasingly insignificant.
  - "As we proceed further and further from the time at which alternative actions are open to us, the events of which either action would be part cause become increasingly dependent on those other circumstances, which are the same, whichever action we adopt. The effects of any individual action seem, after a sufficient space of time, to be found only in trifling modifications spread over a very wide area, whereas its immediate effects consist in some prominent modification of a comparatively narrow area. Since, however, most of the things which have any great importance for good or evil are things of this prominent kind, there may be a probability that after a certain time all the effects of any particular action become so nearly indifferent, that any difference between their value and that of the effects of another action, is very unlikely to outweigh an obvious difference in the value of the immediate effects."
- Not plausible: consider esp. identity-affecting effects.
  - Even for trivial actions, e.g. helping an old lady to cross the road.



Avoiding cluelessness about objective betterness (II): Cancelling

- The cancellation postulate: individual effects (including remote effects) are significant, but unforeseeable effects cancel each other out in the long run (with probability close to 1).
- This is just incorrect: Cf. random walk theory.
  - (The average net effect is zero, but the average *magnitude* of the net effect is nowhere near.)
- Interim conclusion: We have no justification for (NR_o). Thus (CW_o) is true.
- …But so what?...

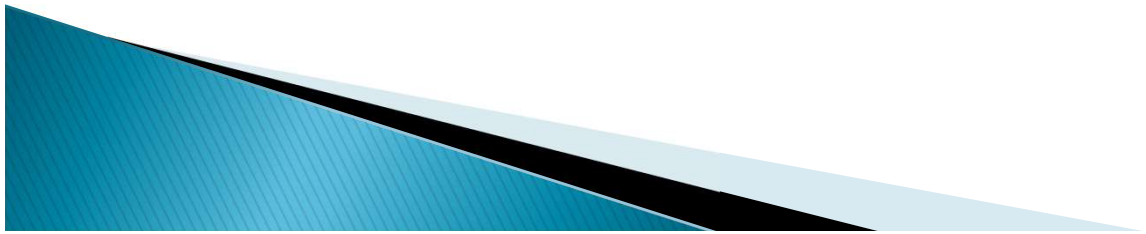


2. Cluelessness about subjective c-betterness

- *Expected* value: The probability-weighted average of possible values.
- **(SB: Criterion of subjective c-betterness)** Act A_1 is *subjectively c-better* than A_2 iff the expected value of the consequences of A_1 is higher than the expected value of the consequences of A_2 (where both expectation values are taken with respect to the agent's credences [or evidential probabilities] at the time of decision).
- **(CW_s: Cluelessness Worry$_s$)** We can never have 'even the faintest idea', for any given pair of acts (A_1, A_2), whether or not A_1 is subjectively c-better than A_2.
  - ?
- **(NR_s (Non-reversal$_s$))** The net effect of taking into account the unforeseeable effects *would not reverse* the subjective c-betterness judgments that we reach based on the foreseeable effects alone.
  - If we are justified in assuming (NR_s), this would rebut the cluelessness worry (CW_s)…

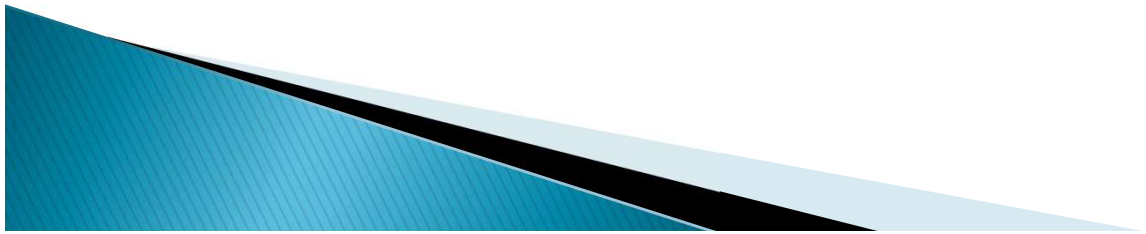
Defending (NR_s)

- For unforeseeable effects E_1, E_2, credence that ($A_1 \Box\!\!\rightarrow E_1$ & $A_2 \Box\!\!\rightarrow E_2$) and credence that ($A_1 \Box\!\!\rightarrow E_2$ & $A_2 \Box\!\!\rightarrow E_1$) should be equal. Therefore unforeseeable effects cancel each other out *in expectation*.
- Thus (NR_s), the analogue of for *subjective* c-betterness, is true.
  - More accurately: Consideration of 'unforeseeable effects' provides no reason for doubting (NR_s) [unlike (NR_o)].
- So we have seen no reason for believing any cluelessness worry about *subjective* c-betterness.



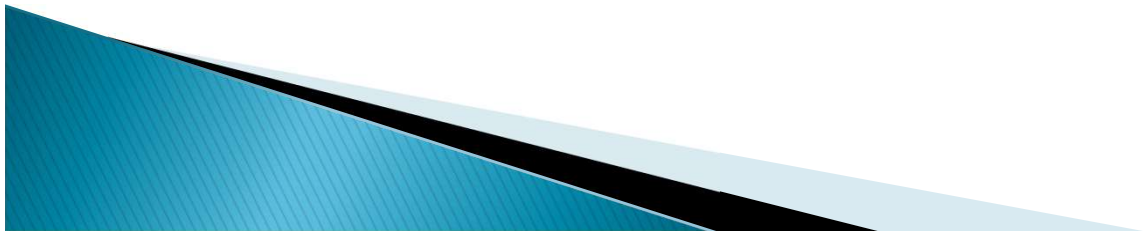
3. Lenman's objection: The Principle of Indifference

- The notorious 'Principle of Indifference' (POI) (roughly): If you have either *no* information or *precisely symmetric* information as to which of n mutually exclusive propositions is true, you are rationally required to assign them equal credence.
- This principle seems right in some cases. (I'm about to flip a coin. The two sides are labelled 'Heads' and 'Tails'. What's your credence that the coin will land Heads?)
- Lenman's objection: The reasoning on the previous slide relies on some form of POI. But the Principle of Indifference leads to 'paradox'…



Paradoxes from the Principle of Indifference: The 'problem of multiple partitions'

- I'm about to draw a book from my shelf. What's your credence that its cover is red? (It's either red or not – so ½? OTOH, it's either red, green, blue, yellow or something else – so 1/5?)
- ('Unnatural partitions'? But similar problems can occur even when each of the partitions in question seems *perfectly natural*.)



Assessing POI

- The current consensus, in response to this problem: POI has initial allure, but is just false.
- Lenman's application of this to the cluelessness argument:
  - …Therefore the above reasoning in defence of (NR_s) ('subjective NR') is unsound.
  - Cluelessness applies to subjective c-betterness, no less than it applies to objective c-betterness.
- Against this:
  - What the 'Problem of Multiple Partitions' shows is that a *fully general* POI is false.
  - But it doesn't show that no *restrictions* of POI are true.
  - And rejecting all indifference reasoning, wholesale, seems to throw out too much baby with the bathwater (both in everyday reasoning, and in science).
  - More optimistic view:
    - There are some true restrictions of POI;
    - We know at least some applications of these when we see them (even if we don't know how to formulate the corresponding restricted principles);
    - The case of interest is one such application-of-a-true-restricted-POI.

4. The 'new problem of cluelessness'

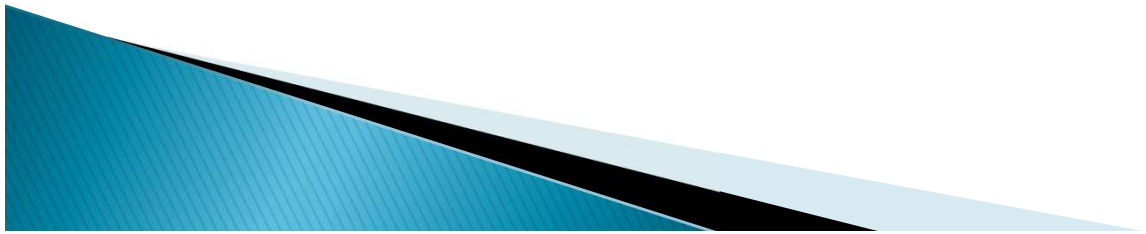
- The cases considered so far are ones in which:
  - There merely *might* be good consequences from choosing A1 over A2, or vice versa;
  - But any reason for favouring some given 'effect-hypothesis' ($A_1 \Box\!\!\rightarrow E_1$ & $A_2 \Box\!\!\rightarrow E_2$) has a precise analog that in precisely the same way favours the rival effect-hypothesis ($A_1 \Box\!\!\rightarrow E_2$ & $A_2 \Box\!\!\rightarrow E_1$);
  - And for that reason, a restricted-POI seemed highly plausible here.
- A different kind of case (the 'new problem'):
  - (NC_1) We have some reasons to think that the unforeseeable consequences of A_1 would be substantially better than those of A_2;
  - (NC2) We have some reasons to think that the unforeseeable consequences of A_2 would be substantially better than those of A_1;
  - (NC3) These reasons are of *quite different characters*, to that it is unclear how to weigh up these reasons against one another.
- No form of POI is at all plausible in *this* type of case.

New cluelessness: An example from 'Effective Altruism'

- Suppose I donate to the Against Malaria Foundation, on the basis of advice that for every $3000 donated I save the life of one child under 5.
  - This corresponds to about $50/QALY – which is cheap!
- But there will be some systematic knock-on effects…
  - Increasing (or decreasing?) population size
    - Either of which could increase or decrease total well-being, via issues of over-/under-population
  - Decreasing political involvement?
  - Affecting wealth levels… and thereby extinction risk
- In aggregate, the *magnitude* is almost-sure to be more than 60 QALYs. But we *really don't know* what its sign is… even in expectation??
- The detailed reasons for thinking that the knock-on effects of saving a child's life might be good are *quite different* from the reasons for thinking that it might be bad (they are *not* merely mirror-image observations that either scenario is *possible*.) And there doesn't seem to be any canonical weighing-up operation.

New cluelessness feels *real* and *genuinely uncomfortable*

- Nobody (except the 'mentally ill') really experiences decision paralysis as a result of the mere fact that moving one's hand/helping an old lady across the road *might* result in e.g. a natural disaster/the birth of an additional dictator.
- In contrast, many would-be 'effective altruists' *do* feel paralysed, in their attempts to do good, by the worry that well-intentioned interventions might turn out *systematically* to make things worse, in ways that are partially foreseen but whose 'probabilities are hard to assess'.
- Three questions, then:
  - What is the right theoretical description of cluelessness?
  - To what extent is it actually true, in cluelessness cases, that consideration of consequences cannot guide moral/practical decision-making or evaluation?
  - What is the source of the phenomenology of deep 'decision discomfort' that attends (genuine) cluelessness cases, for agents who are at least approximately rational?

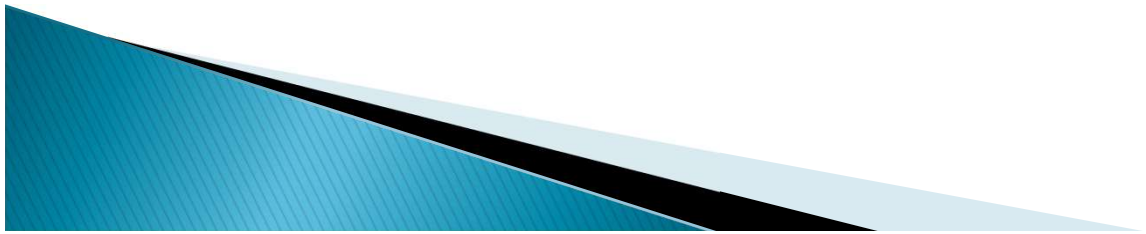


5. The nature of cluelessness

- Not much has been said about *exactly* what cluelessness ('new' or 'old') amounts to.
- The foil: Why isn't it *just* that
  - Each agent has to settle on his credences for the relevant propositions, with or without guidance from a Principle of Indifference or anywhere else;
  - Subjective c-betterness for him will then be determined straightforwardly by those credences, whatever they turn out to be?
  - No obstacle to consequences guiding decision-making/evaluation, via the standard subjective-c-betterness route
  - And no cause for deep discomfort
- Two routes to explore
  - Failure of 'Uniqueness'
  - Imprecise credences

First route to cluelessness: via failure of Uniqueness

- The ‘Uniqueness thesis’: In any given evidential situation, precisely one credence function is rationally permitted.
- Uniqueness very plausibly fails in ‘cluelessness cases’.

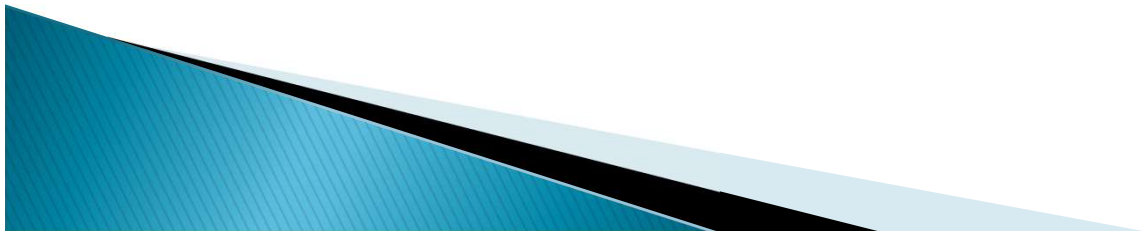


(1) Cluelessness as awareness of uncomfortable arbitrariness? (Some inconclusive thoughts…)

- The agent is rationally required to have some precise credence function.
- Theory gives her no guidance as to which to have (within the permitted class). Her choice must be *arbitrary*.
- But hang on: arbitrary choice is required in 'Buridan's Ass' scenarios too, and in *those* scenarios isn't at all uncomfortable.
  - So there must be more to the present case than arbitriness (if cluelessness-discomfort is indeed real).
- The difference: In the present case, *matters of importance hang on* the 'arbitrary' choice. (?)
- Thus, 'cluelessness as awareness of uncomfortable arbitrariness'?
- Misgivings: But all of the choices here are ratifiable…

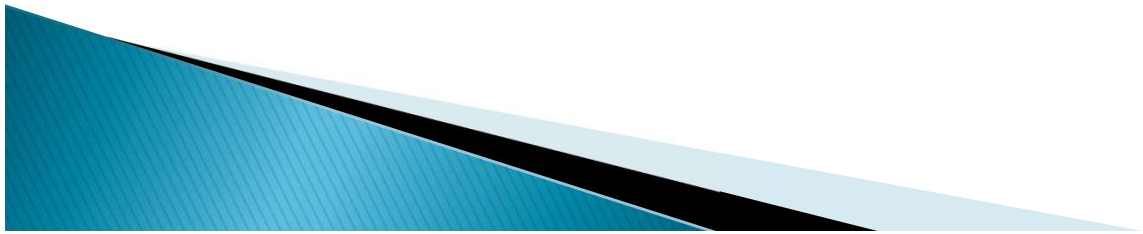
Second family of routes to cluelessness: Via imprecise credences

- Perhaps, in 'clueless' cases, one is *rationally required* to have an *imprecise* credal state.
- Represented by a *set* of probability functions.
  - (Roughly: Interval-valued probabilities, rather than real-valued probabilities.)
- Restricting the discussion henceforth, for tractability: Suppose that *in cases of precise credences*, one is morally required to maximise expected value. ('Subjective consequentialism.')
  - This doesn't immediately tell us what the moral requirements are in *imprecise*-credence cases.
  - One's 'representor' may contain probability functions p_1, p_2, where A_1 maximises expected value with respect to p_1 (resp. p_2). What then?
    - Various approaches: Liberal, Conservative, Maximin EV…
    - (Normally discussed as a topic in the theory of *rationality* rather than morality, but we can piggy-back.)



(2) Cluelessness via imprecise credences + 'Liberal' (max-con) criterion of subjective moral requirement

- Liberal maximising-consequentialist theory of moral requirement under imprecise credences: Act A is morally permitted iff A maximises expected value with respect to some credence function in the agent's representor.
- 'Cluelessness' in the sense that: Both actions A1, A2 are morally permitted, so the agent has to choose arbitrarily?
- But this is just like Buridan's Ass.

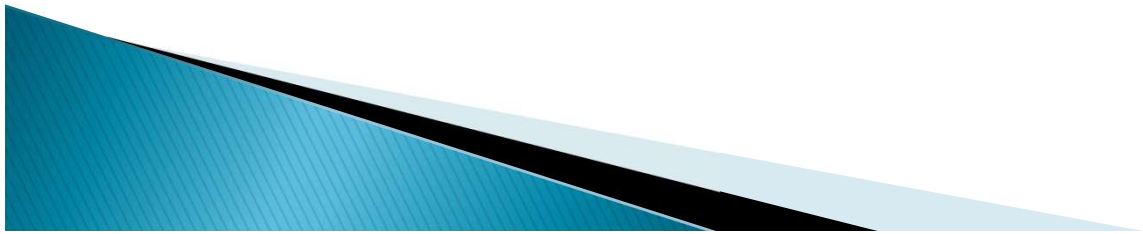


(3) Cluelessness via imprecise credences + 'Conservative' (max-con) criterion of subjective moral requirement

- Conservative maximising-consequentialist theory of moral requirement under imprecise credences: Act A is morally permitted iff A maximises expected value w.r.t. all credence functions in the agent's representor.
- Then 'cluelessness' cases are moral dilemmas (of a new sort, arguably more innocuous than usual).
- But it's unclear why moral dilemma *per se* need lead to cluelessness.
- Again, this case looks relevantly analogous to Buridan's Ass.

(4) Cluelessness via imprecise credences + 'Supervaluational' (max-con) criterion of subjective moral requirement

- Supervaluational maximising-consequentialist theory of moral requirement under imprecise credences:
  - Act A is *determinately morally permitted* iff A maximises expected value w.r.t. all credence functions in the agent's representor.
  - Act A is *determinately morally forbidden* iff A maximises expected value w.r.t. *no* credence function in the agent's representor.
  - Otherwise it is *indeterminate* whether or not A is morally permitted.
- In 'cluelessness' cases, for each A_i, it is indeterminate whether or not A_i is permitted.
- This is uncomfortable, since the agent seeks to match his action to the permissibility facts, yet his *action* must be determinate, and so it can at best be indeterminate whether or not he's successful.
- Note that (unlike the Liberal and Conservative cases) it's not obviously true here that 'all available options have the same moral status'.
  - (Arguably, 'indeterminate permissibility' is not in the relevant sense a moral status.)
- If so, there's perhaps more scope here for the discomfort to amount to a sense of cluelessness.

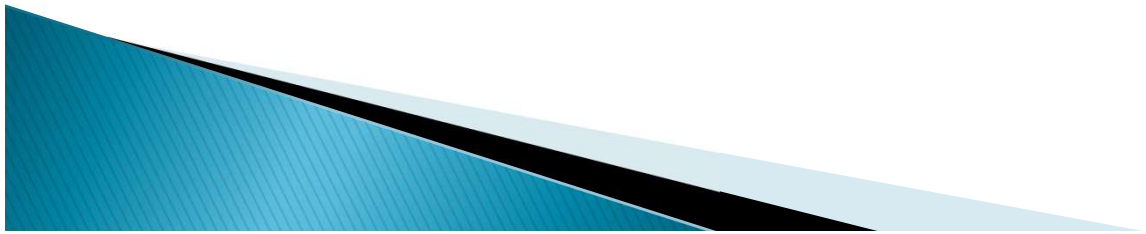


In search of cluelessness: an inconclusive summary so far

- All 4 of the above accounts deliver a sense in which 'consideration of consequences offers no guidance' *within* a cluelessness case. (Which is, however, perhaps not so bad.)
- But in terms of phenomenology: Imprecise credences + Liberal/Conservative permissibility principle seems to deliver at most a very shallow sense of cluelessness.
- So those who feel the pull of *decision discomfort* in (new or old) cluelessness might thereby be drawn to either a Uniqueness-failure account, or an account of imprecise credences together with a Supervaluational permissibility principle, to describe those cases.

7. Mundane cluelessness

- The ‘EA examples’ of new cluelessness feel particularly *vivid*.
- But cases with the structure I outlined are in fact ubiquitous.
  - Choosing a national policy, a career, a diet…
- If cluelessness is no deep phenomenon in these more ‘mundane’ cases, perhaps
  - The Liberal/Conservative approaches are appropriate for the structure I outlined after all.
  - I have misdiagnosed the source of the intuitive cluelessness in the ‘EA examples’.



8. Conclusions

- ‘Old objective cluelessness’ is real, but is no *problem*.
- ‘Old subjective cluelessness’ is not real.
  - Because a relevant restricted Principle of Indifference is acceptable.
- But ‘New cluelessness’ feels(?) real even at the subjective level.
  - At least in some cases (e.g. ‘EA cases’).
- It’s not clear (though) what is the right theoretical description of cluelessness.
  - I have explored some possibilities, involving Uniqueness failure or imprecise credences. (More work to do here.)
- And ‘mundane’ cases (anyway) cast some doubt on my diagnosis of the source of the ‘new cluelessness’ feeling.