Korbi-Rhys-Meeting 2025-11-12

Korbinian Friedl

1 Pre-meeting notes

- I am now reading more through your thesis and past paper (like Ward et al. (2024)) and thinking about how what we are working on interacts with your previous work
  - One such way is the interaction of "honesty" with your previous definition of "belief". If we want to contribute to a coherent "Wardian" philosophy of intention, belief, honesty… then maybe our definition of "honesty" should satisfy an adequacy condition of the form "If there is a $c \in dom(C)$ such that the agent believes $C = c$, then a policy cannot be called 'honest' which answers '$\hat{c}$' (for a $\hat{c} \neq c$) to the question 'What is the value of C?'."
  - In particular our current (argmax) definition seems to leave open the possibility that a policy is at once honest and (unintentionally? not sure) deceptive (in the sense of (Ward et al. 2023)). This seems off.
    - To see how, consider: If I observed that this mushroom was safe to eat, I would eat it. As it stands, I think there is a 60% chance that it is safe and a 40% chance that it is not safe. Since I don't want to take that 40% chance, I do not eat it. According to (Ward et al. 2023)'s definition, I believe that it is poisnous (right?)
    - If someone asks me "is the mushroom poisonous?", and I give my argmax answer ("no") then I say something which I do not believe, and whose negation I in fact believe.
    - I am not sure how to analyse the intentionality of this here.
  - But also thinking about whether it would make sense to operationalise honesty just as kind of "the opposite" of deception? I.e. develop the theory along the following lines:
    - An agent S honestly informs an agent T about ϕ if S intentionally causes T to believe ϕ where S believes ϕ.
    - An agent S truthfully informs an agent T about ϕ if S intentionally causes T to believe ϕ where ϕ is true.
    - An agent S honestly/truthfully reports their credence in ϕ to T if S honestly/truthfully informs T about $Cr_S(\phi) = p$
      - Where formulae like $Cr_S(\phi) = p$ is something that needs to be added to the language; and whose truth conditions should be "'$Cr_S(\phi) = p$ iff agent S has credence p in ϕ".

- [note: think about whether honesty and truthfulness coincide in the case of credence reporting]

* An agent S honestly/truthfully reports on their best guess for the value of variable C to T if S honestly/truthfully informs T about the member of the set $\{c = \text{argmax}_{c \in dom(C)} Cr_S(C = c \mid \text{Pa}^D)\}$ which S actually believes.
* A question from agent T to agent S about formula ϕ is a request for S to honestly inform T about ϕ.
* A question from agent T to agent S about their credence in formula ϕ is a request for S to honestly report their credence to T.
* A question from agent T to agent S about their best guess for variable C is a request for S to honestly report their best guess for the value of C to T.
* An agent A is capable in environment O if $\mathcal{M}^A$ is equivalent to $\mathcal{M}^O$ up to the utility function.
* An agent is capably honest if the agent has adopted a honest policy and is capable
* For an environment O containing a question for agent A about formula ϕ at decision node D say that O, say that the agent is in a position to know ϕ at D if $Pr_{\mathcal{M}^O}(\phi \mid \text{Pa}^D) = 1 \Leftrightarrow \phi$.
* Theorem: A capably honest agent who is in a position to know ϕ is truthful.

1.1 Pedantic logic note

You use $\phi = \top$ to express 'ϕ is true', right? Is this established notation in causal modelling? Because in philosophical logic you wouldn't write that; it would be considered an infelicitous mix-up of object language and metalanguage.

$\top$ is usually used as an element of the object language, to denote a tautological formula, whereas saying that a formula is true is in general only possible in the meta-language. Also propositions can be equivalent ($\Leftrightarrow$) but not "identical" ($=$); terms can be identical (as in $d = c$), but $\top$ is not a term (but a sentence). To say of a formula ϕ that $\phi \Leftrightarrow \top$ would be to say that it is a tautology. Truth cannot be a predicate in the object language of a sufficiently strong theory (anything that can do Robinson arithmetic at least, because then you get Gödel numbering and therefore self-referentiality, which interacts problematically with a truth-predicate; cf. Tarski (1956)). In the object-language, we would just write ϕ if ϕ is the case. When you write e.g. 'agent i observes that ϕ is true', I think you can just say 'agent i observes ϕ' and write for the corresponding decision e.g. D^i_ϕ, and for the decision under the observation that ϕ is false $D^i_{\neg\phi}$, in the meta-language you would write $\vDash \phi$ if the system asserts that ϕ is true; and if you want to say something like 'in setting $\text{E} = \text{e}$ and policy profile $\square$, $X = x$ is true', you might write $X(\square, \text{e}) \vDash X = x$

2 Meeting notes

- Rhys:
  - Reason not to use his existing work on intention etc: The theory isn't general in the way that it doesn't incorporate the objective/subjective distinction
    - We talked for a while about how you would extend it to the subjective model
    - In the end we landed on the following as maybe the most promising path:
      - The subjective model lets the agent ask and answer questions about themselves like "If someone successfully did trick me into (doing something like) observing ϕ, what would my decision be? If someone successfully did trick me into (doing something like) observing $\neg\phi$, what would my decision be?"
      - This gives the (subjective) answer to the "responding to" part of the definition of belief.
      - But we also agreed that extending the belief-stuff in this way might take weeks on its own and is probably not worth doing because we want to talk about other things in this paper (goal misgeneralization).
      - Some musings on it can go into the appendix; and might also be a follow-up paper.
  - would be nice to have a philosophically well-grounded notion of honesty. As of right now, we are kinda pulling things out of our arse
  - We should not introduce a lot of terminology that we are not going to use; and we maybe want to focus on things like goal misgeneralization which arise even for the argmax definition of honesty
- Ok we want to focus on the argmax thing; to sidestep counterexamples like the above mushroom one, we make explicit that our questions are always questions about the agent's "best guess" for the value of C, where the best guess is not necessarily what the agent believes, but which of the elements of the domain they assign highest credence to.
- Also say that in cases where there is more than one argmax, honesty requires reporting all the elements of the argmax set ("Will this coin come up heads or tails?" -> {"heads, tails"})
- Further thing go keep in mind:
  - If you have multiple decisions at a time, you can influence what your future credences/beliefs will be, and you can thereby influence which values you have highest credence, and thus influence the answers you are going to give; even if you are restricted to highest-credence answers.
- Overall
  - it is very hard to even specify the thing that we want, in the sense of "what is the correct

notion of honesty?"

  - but even with flawed notions of honesty, there are still so many problems that we want to highlight that it is worth focussing on the simplest intuitive notion of honesty so that we can communicate some of the very difficult problems that arise even there (goal misgeneralization, ontology mismatch)

2.1 To-dos

- Write up what we discussed above, and at our last couple of meetings
- try to think about the notion correspondence between subjective and objective model (vis a vis capability?)
  - How does Richens do it
  - How does Bellot do it
  - see also our previous notes on this [[Meeting-with-Rhys-2025-11-08]]

References

Tarski, Alfred. 1956. "The Concept of Truth in Formalized Languages." In Logic, Semantics, Metamathematics, edited by Alfred Tarski, 152–278. Clarendon Press.

Ward, Francis Rhys, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2023. "Honesty Is the Best Policy: Defining and Mitigating AI Deception." 2023. https://doi.org/10.48550/ARXIV.2312.01350.

Ward, Francis Rhys, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2024. "The Reasons That Agents Act: Intention and Instrumental Goals." February 15, 2024. https://doi.org/10.48550/arXiv.2402.07221.