Cognitive Science & AI Safety - Journal Club

— computational, machine psychology —ai alignment, risk and safety—

1. Introduction
    - We discuss ***if and if how*** cognitive science has a role to play in AI safety and alignment
    - A significant chunk of the problem of alignment is tied to human cognition and human preferences. Our needs, wants and capacities are the anchor which AI safety wants to operate around, yet the field of AI safety features only a limited discussion of cognitive psychology. Cognitive Science has decades of experimental designs to show for, all of which try to understand the mechanics of black box humans via input and output only.
    - To this end we read papers or book chapters that lie at the intersection between AI alignment and computational cognitive psychology. To get a flavour, see eg. this, or this or this or the reading list we will continue to extend below.
    - We will *not* ask if humans or machines are more intelligent.
    - We will ask how we can study the risks inherent in intelligent systems by dissecting their behavioural capacities into cognitive components or 'cognitive gadgets' (e.g. causal reasoning, active inference, language processing, information-seeking, habitual responses, memory, etc.), to understand how, given their particular constraints (e.g. neuronal refractory periods in brains or data volume in machine learning systems) these cognitive gadgets interact to allow for or constrain risky behavior (e.g. drunk driving in humans or reward-hacking in reinforcement learning agents).
    - This way, we want to assess whether we can theoretically and experimentally build and compare risk profiles of intelligent agents that have different cognitive gadgets and constraints

- Theory is fun and intelligence an interesting topic. But we want to really try and think about whether this line of inquiry helps anyone *solve* problems in AI risk/policy/safety.

Exemplar keywords

- computational cognitive science , casual cognition, robustness, norm adherence,
- ai alignment and safety, comp. psychiatry , deliberation, mechanistic interpretability
- moral psychology/cognition, collective intelligence , trust, reliability, user-design

2. Objectives
    - Being up-to-date with research at the intersection of cog sci and ai safety
    - Collecting research questions that have not been addressed and should be addressed
    - Sharing reading materials
    - Discussing study designs
    - Finding collaborators
    - Addressing the questions
3. Format and Structure
    - 5 min reminder/summary of the paper (i.e. you must have read the paper beforehand to benefit from the discussion)
    - 5 min - attendees raise discussion points - both high level ('why do we care about this paper at all?') and low level ('I didn't understand figure 7.C')
    - We spent most of the hour discussing these, followed by:
    - 10 min zooming out:
        i. what does this mean for AI safety/alignment?
        ii. did we come up with new project ideas that we want to elaborate on?
        iii. is this idea genuinely solving a problem in the world?
    - 5 min: checking in with follow-ups of references we want to send each other, voluntary commitments, homework/questions that we want to investigate for ourselves and admin for meetings

4. Target Audience
    - People who actively work on the interests covered by keywords above
    - Researchers in psychology/ cognitive science/ neuroscience (learning towards computational, but not a requirement)
    - Researchers in AI safety / alignment
5. Schedule and Logistics
    - 1 hour , Friday 5pm, either at Room 10.38, Trajan House, Mill St Oxford OX2 0DJ, EnglandOxford or online
    - There's absolutely no obligation to attend every time. Only spend time on this if it makes sense given your interests and goals. We run the sessions so that they make sense in isolation.
6. Some questions we might want to ask:
    - Can CogSci provide tools and metrics to measure reasoning capacities in algorithmic intelligence (Is GPT really doing X)?
    - Can CogSci tell us anything about how to infer cognition of proprietary black-box algorithms?
    - Can CogSci help us identify the most risky cognitive gadgets (under what constraints and assumptions is and isn't agency a problem?)
    - Can CogSci tell us what makes people less easily fooled by algorithms?
    - Can CogSci help us elicit true preferences?
    - Can CogSci ...
7. Evaluation and Feedback
    - This group is supposed to serve a need and interest, which means that we want to iterate with the format fast. We will drop or intensify these meetings as they serve us and our research: if we find that psychology has not much to say about AI safety and we'd rather spend our time investigating other questions, great! If we find research projects on the basis of our readings and discussion that take up our time, great!

- This means we want to amend the structure and purpose of these meetings as we see fit, so please message the organisers with a low bar to ask for what you want.

8. Contact
    - carla.cremer@queens.ox.ac.uk = this person, can also be reached on Trajan H slack
9. Some readings past and future (feel free to add suggestions)
    - https://github.com/beyretb/AnimalAI-Olympics
    - https://arxiv.org/abs/1604.00289
    - https://link.springer.com/article/10.3758/s13423-020-01825-5
    - https://psyarxiv.com/6dfgk
    - https://global.oup.com/academic/product/natural-general-intelligence-9780192843883?cc=gb&lang=en&
    - doi: 10.3389/fncom.2016.00094
    - https://ought.org/updates/2022-04-06-process
    - https://www.pnas.org/doi/10.1073/pnas.2221180120
    - https://pages.ucsd.edu/~bkbergen/papers/cogsci_2022_nlm_affordances_final.pdf
    - https://link.springer.com/article/10.1007/s43681-022-00188-y
    -