📝 Notes

Nov 13, 2025

David H x Mahault

Invited davidh7057 Mahault Albarracin

Attachments David H x Mahault

David H x Mahault - 2025/11/13 08:25 EST - Recording

Meeting records Transcript Recording Recording 2

Summary

David Hyland and Mahault Albarracin reviewed their presentation for an experiment, addressing the "incredibly slow" simulation speed challenge in Experiment 3 and deciding on a symmetrical presentation structure where David Hyland covers motivation and concepts, and Mahault Albarracin covers the five experiments. Key concepts discussed included path flexibility, resilience, attractor difference, and precision control, which David Hyland and Mahault Albarracin refined, leading to the formulation of three key objectives, including demonstrating the value of their alignment formalization and developing methods to influence attractors. Mahault Albarracin presented five experiments, including testing how coordination emerges from adaptive preferences (Experiment 2), how plasticity promotes recovery after shocks (Experiment 3), and how asymmetric precision control can influence attractors (Experiment 5).

Details

Notes Length: Standard

- **Initial Setup and Planning Discussion** David Hyland and Mahault Albarracin started the meeting to review slides for an experiment, with Mahault Albarracin having an extra slide they planned to adjust visually. David Hyland shared a challenge with Experiment 3, where the necessary long planning horizon for the

agent to backchain made the simulation "incredibly slow," failing to finish running even after 20 minutes (00:00:00). The discussion confirmed a structure where David Hyland would cover the motivation, key concepts, questions, and hypotheses, and Mahault Albarracin would cover the experiments (00:38:24).

- **Review of Presentation Structure and Experiments** The team discussed the flow of the presentation, deciding to keep both the questions and hypotheses, acknowledging that they are rephrasing of the same ideas (00:39:32). David Hyland noted that the presentation felt asymmetrical, and Mahault Albarracin suggested that they might remove Experiment 6 since it was not explicitly addressed in the questions and hypotheses. They decided to proceed with five experiments, which allowed them to divide the slides more symmetrically, including David Hyland handling the synthesis section (00:40:42).
- **Review of Core Concepts and Motivation** David Hyland presented the motivation, highlighting that alignment approaches should address real-world properties such as non-stationarity, diversity of beliefs, bounded rationality, and collective properties, suggesting a focus on "paths and processes" (00:45:31). Key concepts defined included path flexibility, empowerment, resilience, attractor difference, plasticity, planning horizon, and precision control (00:47:15). They also emphasized the importance of embodiment and embeddedness for alignment approaches and introduced the need to better integrate the concept of "common grounds" into the motivation and experiment context (00:52:56).
- **Discussion on Attractors and Power Asymmetries** David Hyland and Mahault Albarracin refined the concept of attractor difference, agreeing that if attractors converge, agents will follow more similar trajectories over time (00:50:11). They also discussed how precision control relates to the ability to shift attractors, leading Mahault Albarracin to suggest phrasing the question to focus on the "power to shift attractors" to better reflect the underlying mechanism of precision control. Mahault Albarracin noted a missing element—the explicit objectives of the work—which they decided should be added to the presentation (01:00:29).
- **Defining Objectives** They formulated three key objectives for their work: to demonstrate the value of their alignment formalization (e.g., the mapping between preferences and belief topologies); to study how path flexibility and empowerment lead to resilience (sustainable alignment); and to develop methods to influence attractors, thereby fostering path flexibility and improvement (01:03:16). David Hyland subsequently integrated these objectives into the presentation structure (01:29:51).

- **Hypotheses and Planning Horizon Concerns** David Hyland reviewed the hypotheses related to their key questions, including the relationship between coordination, communication, path flexibility, and resilience, and how attractor difference and plasticity influence path flexibility (00:54:23) (01:31:15). They discussed the question about planning horizons, with David Hyland expressing concern that long-term planning could make things more brittle if all effort is focused on a specific contingency, raising the issue of "depth versus breadth of planning" (00:57:12).
- **Experimental Plan Overview** Mahault Albarracin presented five experiments, starting in grid worlds but intended to be complex enough for common grounds. Experiment 1 seeks to establish a baseline for coordination benefits, manipulating communication and shared observations (01:05:18) (01:32:43). Experiment 2 aims to show alignment emerges from adaptive preferences by manipulating agent plasticity and update rates to observe convergence speed and joint resilience (01:09:27). Experiment 3, involving periodic shocks, focuses on how individual and shared plasticity affect group recovery time and resilience (01:10:36).
- **Extended Experiments and Synthesis** Experiment 4 connects the "cognitive light cone" or planning horizon extension to system resilience, manipulating the planning horizon's cost and the shock magnitude/frequency, expecting extended horizons to improve adaptation at a metabolic cost, incentivizing group coordination (01:11:52). Experiment 5 seeks to demonstrate the ability to actively influence attractors, with one agent modulating another's precision, anticipating that asymmetric precision control will distort path flexibility and generate dependency (01:13:06). David Hyland concluded with a synthesis outlining future empirical and conceptual questions, including linking findings to empowerment and other alignment metrics, and planning for implementation in common grounds to study embodiment and spatial arrangement effects (01:14:37).
- **Presentation Wrap-up and Next Steps** They made minor adjustments to the presentation, such as attempting to loop an animation and crop a video title (01:16:19). David Hyland performed a run-through of their initial presentation segments, which covered the motivation, objectives, and key concepts (01:28:30). Mahault Albarracin then took over to introduce the first two experiments, stating that coordination leads to better outcomes and that the next step is to show how that arises from their formalism (01:32:43).

- **Aligning Attractors via Adaptive Preferences** Mahault Albarracin presented an experiment designed to show that alignment in agents emerges from adaptive preferences, not just identical goals. The setup involves two misaligned agents where preference update rates and preference plasticity are manipulated, expecting that higher plasticity will lead to faster convergence and higher joint resilience, measured by attractor difference, time to coordination, and long-run efficiency. David Hyland suggested skipping the presentation of the videos demonstrating this experiment (01:33:57).
- **Plasticity and Group Resilience After Shocks** Mahault Albarracin outlined a third experiment aiming to demonstrate that plasticity promotes recovery and adaptation after periodic resource relocation and noise shocks in a multi-agent grit world (01:33:57). This experiment manipulates high individual or group plasticity, social posterior sharing (belief sharing), and redundancy, with the expectation that learning from errors and sharing corrections will lead to faster, deeper recovery and turn resilience into a group-level property. Albarracin also mentioned a short experiment based on work by Rudy Pitia Jane and Timber Verbalen that showed belief sharing and theory of mind allowed agents to recover more quickly from shocks (01:35:04).
- **Group Planning and Path Flexibility** Mahault Albarracin proposed an experiment to show that planning horizons must be extended to maintain path flexibility and that group planning yields greater rewards because it reduces the individual metabolic cost of planning. The setup involves agents with variable planning horizons facing unexpected resource shocks, manipulating the cost of horizon expansion, shock magnitude, and frequency, with the expectation that extended horizons yield better adaptation but are more efficiently managed as a group. Finally, Albarracin proposed testing whether power in multi-agent inference arises from prediction control, showing that one agent can modulate another's sensory or policy precision via environmental actions, potentially leading to agent alignment without losing path flexibility (01:36:21).
- **Synthesis and Next Steps** David Hyland provided a synthesis of the preliminary ideas, emphasizing that significant empirical work remains. Hyland noted the need to study the connections between past flexibility, resilience, and energy efficiency, and conceptually link findings to empowerment, shared intentionality, and other alignment metrics. For implementation, they plan to build a prototype in common grounds to incorporate embodied aspects and show how spatial properties affect path flexibility and resilience, with expected deliverables

including theoretical formalizations, mathematical results, empirical analyses, and data visualizations (01:37:38).

Suggested next steps

No suggested next steps were found for this meeting.

📖 Transcript

Nov 13, 2025

David H x Mahault - Transcript

00:00:00

David Hyland: Hello. How are you?
Mahault Albarracin: I'm good. I'm just uh working on something to um make new experiment three a little more interesting. Uh but ultimately have the slides so we should be fine.
David Hyland: Sure.
Mahault Albarracin: Um, if you want to go run through them once and then we can maybe um, you know, record whatever it is you want.
David Hyland: Um, yeah, I was trying to, but um it didn't seem to end up working.
Mahault Albarracin: Did you manage to um, produce an experiment? I know you were trying to. That's okay.
David Hyland: Sorry. Um, I think it's I guess one of the issues was that the in order for the agent to figure out which direction to try and move in, it has to plan like the planning horizon has to be sufficiently long so that it can actually backchain. But then that ended up making the thing like incredibly slow. like there was something but it just never finished running even after like 20 minutes. So So Oh, okay.

00:36:48

David Hyland: I see you have here. Um yeah. Okay. Should we just quickly talk through it and then um Which one is this?
Mahault Albarracin: Yes, it's just that there's an extra slide in there that's the same slide where I was waiting to slot in something, but uh just just below. Yeah, this one. So, you can ignore that one or make it for now. Um, I'll just skip it. I was going to change the visual. Done.
David Hyland: Okay.
Mahault Albarracin: Okay.
David Hyland: Um, should we let's see
Mahault Albarracin: Oh, f***. I need to open it in a different browser. Just give me one second. Here you go. m*********. I don't have a Just sec. Share with M. My computer is

incredibly slow this morning and I don't know why. Like it's not moving. It's just like stuck and then I have to wait like three minutes and be like, "All right, you going to unstick?" Like, and I have no clue why.

00:38:24

Mahault Albarracin: Like, it's just the way it is. Do you know that song? No, you're too young. Okay, that's just the way it is.
David Hyland: Which which is this. Don't think so.
Mahault Albarracin: It's a very famous song.
David Hyland: Okay. Um, yes. So, I guess there's a side two and three. Well, three is like a duplicate of two.
Mahault Albarracin: Yeah. Well, so that's what I um I put one on um Wait, what? Oh, did you duplicate it?
David Hyland: Yes, I duplicated and made some changes or like rearranged some stuff.
Mahault Albarracin: Okay. Do you want to remove this one?
David Hyland: I'm not sure. Sorry.
Mahault Albarracin: Do you want to remove the first one?
David Hyland: Sure. We can just hide it. Um so I think in terms of structure um since you focused on the um experiments do you want to talk through those and then I'd be happy to talk through like the intro or or how do you want to structure it?

00:39:32

Mahault Albarracin: Yeah, that works for me. If you want to do that, you can you can talk through the motivation and the key concepts and the questions and the hypothesis and then I'll just talk through the experiments.
David Hyland: Okay. Um, do we want to talk about the questions or just focus or just present them as a hypothesis because it's basically just a rephrasing.
Mahault Albarracin: that it is exactly it is exactly that. But um you can if you prefer not to have questions you can just say we have hypothesis you know it's just that that means I have a very large chunk of the presentation.
David Hyland: Okay, maybe maybe I'll just rephrase some of the hypothesis. Or I guess we can we can we can do the questions and then say okay here's here's like our well we we we've like introduced the questions and say here are our hypothesis related to these

questions.
Mahault Albarracin: Yeah. U but you can you can talk to them at the same talk about them at the same time and combine the slides if you want so that you talk about both.

00:40:42

Mahault Albarracin: But I I'm still sort of worried that it's um very asymmetrical and also we have never run through it and I kept experiment uh effectively wait right wait three this was four I didn't change
David Hyland: Yeah. Um
Mahault Albarracin: and this was five now um wait five. Yeah. And then this would have been six now. But I'm I'm not sure we want to keep experiment six anymore. I'm happy to keep it, but um I'm not sure we want to.
David Hyland: Yeah, maybe because we don't really address it in the questions and hypothesis.
Mahault Albarracin: That's it. That's it. It's kind of implied in the sense that it's, you know, it's function of whether resilience help.
David Hyland: I think yeah.
Mahault Albarracin: But we don't have to do it. We can we can skip it.
David Hyland: Yeah, I think it Yeah, sure.
Mahault Albarracin: Um so that means we still have five experiments. So you can do the synthesis then which means you still have two three four five slides and I have about the same number.

00:41:59

Mahault Albarracin: Does that work for you?
David Hyland: Yeah, sounds good.
Mahault Albarracin: Okay.
David Hyland: Um, so I think just change these arrows cuz they're a bit hard to see.
Mahault Albarracin: Do you want to present the screen? Share screen. Sorry.
David Hyland: Um, these pink arrows are a little bit hard to see on the lav Um, okay. Should we just do like a quick run through and then we can just um actually hold.
Mahault Albarracin: Yep.
David Hyland: Do we want to combine questions and hypotheses? How about I talk

through the hypothesis on the question slides as I put those in the notes.
Mahault Albarracin: Yeah. Yeah, absolutely.
David Hyland: Just the notes. Whoa.
Mahault Albarracin: But in any case, the run through starts with you, right? So I'm just And also I'll let you share the the slides first because it's again your Okay.
David Hyland: Okay, there aren't any notes for some reason. Prisoner review. What is that?
Mahault Albarracin: Yeah.
David Hyland: Okay, let's see. Does it change automatically?

00:45:31

David Hyland: Oh, nice. Okay, cool. Hello everyone. Um today Ma and I are going to be talking briefly about some ideas we have around path flexibility and collective alignment. So just to start as motivation, why do we care about path flexibility and why might it be an important concept for alignment? Well, our starting point is that approaches to alignment should address some key properties of the real world, which include non-stationarity of the environment, non-stationity of and diversity of human beliefs and preferences, the fact that all agents are bound and irrational and can fail to execute their plans perfectly. And there are collective and system level properties that should be taken into account when we're talking about alignment. These features suggest that we focus on the paths and processes involved in both individual and collective plans. So this means that we think about the alignment of not just beliefs and preferences but the the ways in which they update. Uh it also involves the ability to efficiently adapt and recover when environments or goals shift and also the maintenance of flexibility and optionality over time rather than rigidity and narrowness.

00:47:15

David Hyland: And so a simple example of this is in the classic lava world scenario where an agent is trying to reach a goal, but there is lava on either side of this walkway. And intuitively the agents should spend a little bit of extra effort to traverse the middle path because in the case that they end up slipping there are fewer disastrous consequences which they are unable to recover from. So some key concepts that we think are important to define and study and address. Uh obviously the first one is path

flexibility which relates to the quantity and ease of shifting between viable trajectories for an agent or a group. Uh empowerment is roughly the degree of influence that an agent can exert on its environment. Resilience is the ability to recover from and adapt to adversity through mechanisms of redundancy and degeneracy. Attractor difference roughly measures the distance between age and stable belief or preference trajectories. Plasticity here we take to be the degree to which beliefs, preferences or policies can update in light of prediction errors in changing circumstances. or in other words, an agent's responsiveness to changing environmental circumstances.

00:48:51

David Hyland: And planning horizon is related to the cognitive light cone, which is essentially the effective temporal depth of active inference. How far into the future is an agent able to plan, make predictions, and care about. And finally, precision control, which is the capacity to modulate uncertainty related to sensory policy or belief precision. And so, um, when when should I bring in this picture?
Mahault Albarracin: Uh so this is how uh attractor difference and implicity. So you want to talk about it during attractor difference but you don't have to remember it's fine.
David Hyland: Yeah. Okay. So, distance between agents. Yeah. Um, do we need this bottom visualization here? Because I feel like Yeah, it's kind of already captured.
Mahault Albarracin: Not necessarily. It's the same thing. So, you can It's It was just meant to show that there's, you know, different attractors in the video. It was just meant to show what happens if you bring them together. Um, but you can you can you can remove the um the picture at the bottom and maybe make the video auto play when you move the

00:50:11

David Hyland: Yes, that's a good idea. This here. How do we do this? Animate. Okay. Play automatically. Let's try that. I think maybe click is better because we we don't talk about it until start. tractor difference, the distance between agent stable of preference of relief trajectories and crucially if these attractors um evolve over time. What am I going to say?
Mahault Albarracin: What?
David Hyland: If these attractors converge

Mahault Albarracin: The idea is that you you you reduce the distance between the attractors. But if you don't like that one, use the other one.
David Hyland: Oh, I mean I like this. So I'm just trying to figure out how do we if the if the attractors converge then the agents converge.
Mahault Albarracin: Okay. Yeah. I mean, that's the idea.
David Hyland: Um, if the attractors converge, then agents will tend to follow more similar trajectories over time. Something Okay.
Mahault Albarracin: Although Yeah. No. At least the Yeah.
David Hyland: What are you going to say?

00:52:56

Mahault Albarracin: No, no, no. It's fine. That's fine. This is correct.
David Hyland: Okay. Okay. So some questions that we what is it I think we talked about at the very end didn't we yeah in the synthesis
Mahault Albarracin: Wait. You know, one thing we still haven't actually done at all anywhere that we were supposed to and we didn't we didn't talk about common grounds like at all. Do we where we should probably we should probably motivate common grounds a little better.
David Hyland: I mean yes implementation Yeah.
Mahault Albarracin: Like if you wanted to put something into motivation, I think that would be good.
David Hyland: Okay.
Mahault Albarracin: and and explain why none of the experiments look like they're in common grounds because we're gonna I mean I'm going to explain I think maybe that's something I should put in there.
David Hyland: Okay.
Mahault Albarracin: Um that for now all these experiments um are in grid worlds but they are meant to then be um made a little more complex to fit common ground.
David Hyland: Okay, maybe we can say something about like the importance of embodiment and embeddedness within an environment for the approaches to alignment.

00:54:23

Mahault Albarracin: Okay. Yeah. Yeah. Great idea. Oops.
David Hyland: Adam, what the f*** is embodiment and embeddedness within the environment?
Mahault Albarracin: We got
David Hyland: Yeah. Okay. Oh s***. Okay. Um, so some questions uh that we think are important to understand and address. First of all, how do coordination and communication contribute towards path flexibility and systemwide resilience? Here we hypothesize that in general imp greater communication and coordination will lead to more path flexibility and increase the capacity for systemwide resilience. Um second question is how does do do higher attraction difference and low plasticity lead to lower collective pl flexibility? Uh we hypothesize that this is the case. Question three is how is plasticity related to the ability to recover or adapt after shocks? and we hypothesize that this generally promotes the ability to recover and adapt. Question four is related to whether extended planning horizons improve resilience and are there diminishing returns and to what degree? Um, I'm just Yeah, I guess this one I had like a hard time like thinking

00:57:12

Mahault Albarracin: Do you have to skip the slides? I can't do it for you. Are you trying to think of like What?
David Hyland: about like Because I don't know, intuitively it's like if you make a long-term plan and then you put all your you sort of put all your eggs in that basket. um could that not potentially make things more brittle or less resilient because you'd sort of like you know planned for this very specific contingency and then and then if things go wrong suddenly you don't know what to do.
Mahault Albarracin: Yes.
David Hyland: So it feels like feels like there's something here to do with like the depth versus breadth of planning.
Mahault Albarracin: Yes. But I Yeah, I thought that was the whole point of the path flexibility empowerment.
David Hyland: Yeah.
Mahault Albarracin: You know, that's one thing we're not actually testing here.
David Hyland: Mhm.
Mahault Albarracin: We have past flexibility. We have attractor difference that leads to

common or at least um at least mapping to to to similarity. Um but then we we're missing the empowerment bit which I think we had in the key concepts.

00:58:44

Mahault Albarracin: No. Yeah.
David Hyland: We we mentioned empowerment in the key concepts.
Mahault Albarracin: But we don't actually have an experiment to that end.
David Hyland: Yeah.
Mahault Albarracin: Should we add Yeah.
David Hyland: I mean I mean in the synthesis we talk about oh yeah we should understand how it relates to empowerment better which I think it's fine to leave it as an open question for now I suppose.
Mahault Albarracin: Okay. But so that's the breath question basically.
David Hyland: Yeah. Okay. Um, so maybe I'll just rephrase this by say how do breading horizons effect. Do we want to talk about resilience or past flexibility or both?
Mahault Albarracin: Um so resilience is your ability to return from a shock. So you can call it path flexibility if you wish. Um but ultimately
David Hyland: No. Yeah. Okay. I think this is resilience. Um. Yeah. Okay. And then how does position asymmetries and precision control influence biodnamics? Um so our hypothesis here is that agents that are able to exert greater precision control or who have higher precision are able to gain maintain and exert more power.

01:00:29

David Hyland: All right.
Mahault Albarracin: Yeah. But I mean so the the idea of the power is the the ability to um to shift the attractors basically that the whole point of that was to bring like okay well all that's is well and good but so what the power symmetry and precision control was how do you ultimately bring in the ability to change those attractors.
David Hyland: Right.
Mahault Albarracin: Um so that's Do you want to make that clearer in the questions?
David Hyland: Right. Mhm. Yeah.
Mahault Albarracin: Okay.
David Hyland: So, how do we how do precision symptoms precision control influence?

How do you think we should phase them?
Mahault Albarracin: The power to shift attractors.
David Hyland: Thanks.
Mahault Albarracin: So I think one thing that we've got missing I guess still sorry is we don't really say anything about our actual objectives I don't think um we say here's the motivation here are the questions and concepts we don't
David Hyland: All right.
Mahault Albarracin: really say why what we're trying to achieve right so the objective ives of this should be maybe above the questions.

01:02:04

David Hyland: Or maybe even in the motivation.
Mahault Albarracin: Yeah, sure. But add yourself a slide like don't pack it all in one spot. But say with this we hope to do to show like three things. One is what are the breaking points? Two is what are ideal scenarios and three how to get to these ideal scenarios. And and you can get a little more you know um No, I think you should put it on the slides.
David Hyland: Oh, should I should I just like say that or do we want to put that on the slides?
Mahault Albarracin: I think it should be clear with the objectives.
David Hyland: Okay. Um, where do we probably Yes.
Mahault Albarracin: Do you have another half hour to go? Okay, great.
David Hyland: Um, magnets.
Mahault Albarracin: Also, you've got I'll I'll fix it. Don't worry. Yeah. Okay.
David Hyland: Aent magnets. Um, okay. Where do I put this? Is it going to be two crowns here?
Mahault Albarracin: Uh yeah, I think you really should have an extra slide.

01:03:16

Mahault Albarracin: Call it objectives. The so so basically the motivation is the why but the objectives is the what. Um and the experiments are the how up
David Hyland: Okay. So, um identify factors that how distinguish
Mahault Albarracin: Uh okay. So um I think we should say objective one is show what

can break. Right. So why um path no objective one is show that our metrics surrounding alignment have value. like for example the mapping between topologies of preferences and beliefs. So that's the first thing we want to do. Uh the second thing is validate that um this the idea towards alignment that we have i.e. path um path flexibility uh empowerment and resilience or no the path flexibility and path empowerment ultimately lead to group resilience which is you know sustainable alignment. Um and then that we have the means to show how we can affect a system such that it leads in that direction. Um that it's not just chance, you know, that we can actually create agents that have the ability to shape the environment or other agents beliefs such that they follow such a desired alignment.

01:05:18

David Hyland: So developed methods for improving. Where's it? Um, how do you want to phrase this?
Mahault Albarracin: I mean develop methods for what? For which part?
David Hyland: What was the third the third thing that you said?
Mahault Albarracin: Oh, um to um influence the attractors basically. to foster path flexibility and improvement.
David Hyland: And then what was the first one again? Show the value of
Mahault Albarracin: Uh the value of our alignment formalization. Sorry.
David Hyland: Is Um, a bit more. Demonstrate the value of alignment formalization. Study how path flexibility and empowerment lead to resilience and develop methods to foster path flexibility concepts questions. Okay. And then do you want to quickly go through your part?
Mahault Albarracin: Mhm. So, in order to test these hypotheses, um, we're going to start with experiments in grid worlds, but they're meant to be made a little more complex for common ground. So, keep that in mind. The first experiment is meant to test the coordination baseline, i.e., Are we even able to find benefits um for coordination to contrast with um alignment implicity manipulation?

01:09:27

Mahault Albarracin: So we're going to have two agents, two resources and they have optional communication at a cost. And so here we're going to have them either

communicate or not and have shared or private observations. And therefore we're going to measure how well they can cooperate, how much they cooperate, how much energy they expend and whether they can return to a specific path given that coordination. Um the expectation is that we will see increased communication uh also increases path flexibility and ultimately shared reward. So they'll have the ability to stay relatively uh coordinated without getting too far away from each other and maintaining a group path.
David Hyland: What?
Mahault Albarracin: In experiment number two, we're trying to show that alignment emerges from adaptive preferences such that they're not just sticking to one thing, but they have the ability to again shift path and therefore shift attractors. So you'll have two different two misaligned agents. They have variable preferences and they have variable update rates. So one of them will update faster than the other. Um we want to manipulate how plastic they are.

01:10:36

Mahault Albarracin: So one of them will have high plasticity and low plasticity. And so what we're going to see is um with attractor difference how long it takes them to coordinates and how efficient they are over the long run. We expect that a high plasticity means a faster conversions and ultimately higher joint resilience. So the ability to um find a better path. In the third experiment, we're going to show that um plasticity promotes uh recovery after shocks specifically. So, we'll have multi multiple agents in a grid world, but ultimately in common grounds with some periodic shocks, and the agents have to learn from prediction errors. Uh we're going to try to see whether low versus high individual plasticity uh impacts the outcome. uh we're going to show that um the fact that they're able to share information and the redundancy versus the no redundancy here, meaning they have very similar models or very different models or and therefore the ability to um get to the similar outcomes. Ideally, what we're hoping for is to show that they have higher recovery time, a lower prediction error, and uh a better resilience index for the group.

01:11:52

David Hyland: That's the
Mahault Albarracin: We're hoping that the learning from errors together means that um

they can share those corrections and create faster, deeper recovery, which turns into a resilience at the group level. So uh if you look at the next one, it's just a small experiment run on top of an experiment that was made by uh Rudy Pitia Jane Jane Pitia um where the agents were able to share beliefs and um perform theory of mind and therefore coordinate. But if you add a shock, how well are they able to recover? So it's just a tiny bit of experiment. Then in experiment four, we're going to try to uh connect the cognitive light cone extension to system resilience and cooperative foresight. So essentially we want to show how a planning horizon affects their ability to re to um return from shocks. So agents here will have variable planning horizons and face unexpected resource shocks. And so the cost of the horizon um we're going to manipulate the cost of the horizon expansion and the magnitude and frequency of the shocks.

01:13:06

Mahault Albarracin: And so here um we're hoping to measure the recovery time, the performance degradation and the path survival rate, i.e. um how well can they maintain path flexibility. And so we're hoping that the extended horizons will yield better adaptation but ultimately at a metabolic cost which incentivizes group coordination. In experiment five um we want to show that given that we've shown the value of mapping to similar attractors whilst maintaining um path flexibility. We want to show that we can actually affect the attractors that it's not just um um a potential outcome. And so we want to have one agent control the attractor of the other. Uh so here one agent can modulate the other's sensory uh or policy precision via environmental actions. And so what is able to manipulate is the control strength and the target precision type. And so here we're going to measure induced policy shifts. cooperation rates and conflict persistence. So, um how well are they actually able to coordinate through these uh through this empowerment? And so, we're hoping that asymmetric precision control distorts the path flexibility, which means you have one agent that's able to shift where the agent goes um and ultimately generates dependency or domination from one of the agents.

01:14:37

Mahault Albarracin: By domination, we really just mean the ability to control the

attractors.
David Hyland: Yep. And so just a synthesis and next steps for uh this research program. Um there are a bunch of empirical questions. Uh so we can study the connections between path flexibility, resilience and energy efficiently in more detail. Uh there are several conceptual questions including um how we link our findings to empowerment. Uh shared intent in intentionality and other alignment metrics have been proposed in literature. And uh further implementation of this beyond grid worlds would be to prototype this in common grounds to bring in questions of how embodiment and spatial arrangements of agents affects questions of path flexibility and resilience. And ultimately the deliverables that we hope to achieve are a theoretical formalization of the concepts proposed here. Uh some mathematical results related to the different quantities that we propose. Some empirical analyses building on top of the basic experiments that we have demonstrated here and some further data visualizations. Thank you.
Mahault Albarracin: So what do you feel like?

01:16:19

David Hyland: Um in what sense?
Mahault Albarracin: Well, do you feel like something needs to be improved? where we need to make changes before recording because we need to record
David Hyland: Yeah, I think I just need to crop this title. Are you Can we crop this title out because it's kind of cut up
Mahault Albarracin: What?
David Hyland: the title of this video? Can we crop it out?
Mahault Albarracin: Oh, sure.
David Hyland: How do we do that?
Mahault Albarracin: Do you not know? Oh, I don't think you can.
David Hyland: No. Oh, I know.
Mahault Albarracin: Nope.
David Hyland: I know. I know what we can do.
Mahault Albarracin: Okay.
David Hyland: We can insert the box. Yeah. Do we want to do we want these animations to just play on a loop?
Mahault Albarracin: ideally. Yeah.
David Hyland: Let's see. How do I do I get this. Don't know. Doesn't seem like I can loop it. Okay. Well, we'll try to time it. Is there anything else that you wanted to change?

01:19:06

Mahault Albarracin: Um, no, not yet.
David Hyland: Okay. Do you want to give it a go?
Mahault Albarracin: Um, no. Uh, oh, sorry, sorry, sorry. My bad. I I mean, not in the presentation. I want to change something, but there's a few things I want to do before we we give it a go.
David Hyland: Okay.
Mahault Albarracin: at the moment.
David Hyland: Okay.
Mahault Albarracin: Okay, I wrote my stuff down.
David Hyland: Okay, let's go.
Mahault Albarracin: Wait, I'm about to record something. Please don't come back in. Don't come back here. I'm about to record something. Okay. Anytime.
David Hyland: Okay, now it's getting very slow. Oh gosh, where are my notes? Okay, this is weird. Sudden suddenly unhappy. Okay. What do you think? Hello. So, Ma and I are going to be talking briefly about some ideas that we've had on path flexibility and collective alignment. So, I guess to start off with some motivation, why do we care about path flexibility?

01:27:38

David Hyland: Um our starting point is that uh a lot of approaches to alignment should really address with sorry can we

Session ended after 01:27:49

01:28:30

David Hyland: Hi. So, Mao and I are going to be uh presenting some thoughts and ideas that we have on path flexibility and collective alignment. So, to start off with some motivation, why path flexibility and what is it? Our starting point is essentially that

approaches to alignment should address some key properties of the real world which include non-stationity of the environment, non-stationarity and diversity of human beliefs and preferences, bounded rationality and imperfect execution of plans, collective and system level properties and also the embodiment embodiment and embeddedness of agents within their environments. And so to address some of these issues, we suggest that we focus on paths and processes. And these things would include the alignment of beliefs and preferences in terms of how they update and not just what they are. The ability to efficiently adapt and recover when environments or goals shift. and also the maintenance and flexibility and optionality over time rather than rigidity and narrowness.

01:29:51

David Hyland: And so a simple example of this is a lava world example where the agent has to go to its goal and traverse this pathway where there's lava on either side. And the basic intuition here is that the agent should expend a little bit of extra effort to traverse the middle path because this provides them more flexibility and resilience in case they accidentally slip um to either side. So some of the key objectives that we hope to achieve within our work are firstly to demonstrate the value of our alignment formalism. Secondly, to study how past flexibility and empowerment lead to resilience and finally to develop methods to promote past flexibility and empowerment. So some key concepts that we find useful to study and address and define. Well, first of all, there is past flexibility which we refer to as the quantity and ease of shifting between viable trajectories for an agent or group. Then there is empowerment which roughly relates to the degree of influence that an agent can exert on its environment. There's resilience which is the ability to recover from and adapt to adversity through mechanisms of redundancy and degeneracy.

01:31:15

David Hyland: And there's concept of attractor difference which is roughly the distance between agents stable belief or preference trajectories. And the key idea here is that the closer aligned agents attractors are the more similar their trajectories will be. And there is plasticity which is roughly the degree to which beliefs, preferences or policies can update in light of prediction errors and changing circumstances. There is planning

horizon which is also related to the cognitive light cone. Uh which essentially measures the effective temporal depth and breadth of active inference or planning into the future. And finally, precision control, which is the capacity to modulate uncertainty uh including sensory policy or belief precision. So some key questions that we want to address. First of all, how do coordination and communication contribute towards path flexibility and systemwide resilience? Here we hypothesize that these things come hand in hand. Uh second question, how do do high attractive difference and low plasticity lead to lower collective path flexibility? Thirdly, how is plasticity related to the ability to recover and adapt after shocks?

01:32:43

David Hyland: Um fourthly, how do the breadth and depth of planning horizons affect resilience and other diminishing returns? And finally, how do precision asymmetries and precision control influence the power to shift attractors? And so with that, I'll pass on to M to discuss some of the experiments.
Mahault Albarracin: Thanks. So for the experiments for now, they're explained in grid worlds, but they're meant to then be made a little more complex to fit the common grounds. So the first experiment is meant to establish a baseline of coordination benefits, i.e. is there even a point to coordinate um and align and so here we have a setup with two agents and two resources with optional communication at a cost and we're hoping to manipulate the communication and the shared observations um to show and to measure whether there's a higher cooperation rates um to measure the expendit the energy expenditure and the path returnability i.e. uh we expect to show that communication increases path flexibility and shared reward. In experiment two, um now that we've shown coordination leads to better outcomes, we must show how it might arise from our chosen formalism.

01:33:57

Mahault Albarracin: So aligning attractors and so we hope to demonstrate that alignment emerges from adaptive preferences rather than just identical goals. So here we have two misaligned agents with variable preference update rates. we're going to manipulate uh the plicity of their preference. And so ultimately we measure the attractor difference, the time to coordination and efficiency over the long run. And we expect that

higher plasticity means faster convergence and ultimately higher joint resilience. I think the videos are struggling to run. So do you want me to wait a bit?
David Hyland: Yeah, we can skip this.
Mahault Albarracin: All right. And so in experiment three um we're trying to demonstrate that plasticity promotes recovery and adaptation after shocks. And so we want to show the mechanisms that underly the plasticity in question i.e. what do the agents learn and from where. So we hope to show that it leads to higher group resilience and path flexibility which means recovering from shock and continuing to thrive. So the setup here is multi-agent grit world with periodic shocks.

01:35:04

Mahault Albarracin: Uh so resource relocation and noise and the a agents learn online from prediction error. Uh we're going to manipulate high individual or group plasticity social posterior sharing. So belief sharing and redundancy versus no redundancy i.e. whether agents can do the same thing more or less or do they have more uh degeneracy. For example, we're going to measure the recovery time, the prediction error, the free energy slope, and the group resilience index, i.e., are they more resilient as a group or not? We expect to show that learning from errors and sharing those corrections creates faster, deeper recovery and turn resilient into a group level property. In the next slide, um, we ran a short experiment on top of a paper that Rudy Pitia Jane and, uh, Timber Verbalen and the rest of their team at Versus ran on theory of mind, uh, with, uh, two resources. And basically, we were able to show that, um, through theory of mind and belief sharing, the agents are more able to recover quickly from their shock. In experiment four, um, we want to show that we need to think in terms of paths and extend the planning horizons such that we can maintain path flexibility over time.

01:36:21

Mahault Albarracin: We also want to show that planning as a group yields more rewards because planning is metabolically costly. So doing it as a group reduces the individual cost. So the setup would be agents with variable planning horizons facing unexpected resource shocks. We would manipulate the cost of horizon expansion. uh the shock um magnitude and frequency and measure the recovery time, the performance degradation and the path survival rate. And so we expect that extended horizons yield better

adaptation but at a metabolic cost. So it's better to uh do it as a group. And finally, we want to show that these are not just chance outcomes. We can actually manipulate paths and lead to alignment. So here we test whether power in multi-agent inference arises from prediction control rather than just force. And so we can show that one agent can modulate another sensory or policy precision via environmental actions. So the manipulation here is the control strength and the target precision type. And we would measure induced policy shifts and cooperation rates uh to show that ultimately we have um the ability to change the attractors for the agents to lead them to potential uh similar paths without necessarily losing path flexibility.

01:37:38

Mahault Albarracin: And so here we have uh an expectation of asymmetric path precision um distorts the path flexibility of one of the agents which generates a dependency or a domination here and basically just control of one agent of the over the other.
David Hyland: Great. And so just as a synthesis and some next steps, obviously these are very preliminary ideas and um there's a lot of empirical work that remains to be done. In particular, we want to study the connections between past flexibility, resilience, and energy efficiency in more detail. Uh there are a lot of conceptual questions. Uh for example, linking our findings to empowerment, particularly on past flexibility, um shared intentionality and other alignment metrics proposed in literature. uh for implementation. In order to bring in the embodied aspect, uh we want to build a prototype in common grounds with some basic features to demonstrate how um spatial properties and embod embodiment affect questions of path flexibility and resilience. And finally, the deliverables we hope to achieve include theoretical formalizations of the concepts that we discussed and introduced and some mathematical results relating them to each other. Uh further empirical analyses both in grid worlds and in common grounds and then some data visualizations. Thank you.

Transcription ended after 01:39:22