# Bookkeeping

First, thank you for your participation in today’s study.

My name is <<insert name>>, and I’ll be helped by <<indicate them and name them>>.

Before we go any further, take your phones out, turn them all the way off, and then put them back in your pockets.

I’ll wait.

*(Turn your own phone off, demonstrating what your expectation is, and wait for them to do this)*

Today, you’ll be evaluating a single artificial intelligence throughout this study, using a tool we’ve built.

In order to make you all familiar with this tool, we’ll have a tutorial for the first little while before tackling the main task.

Please do not talk to one another throughout the study.

We want to see how each of you as individuals perform in the task.

I’ll be reading from this script to be consistent in the information I provide you and the other people taking part in this study.

You’ll be referred to by your participant number for this study to answer questions, so go ahead and find that number now on your desk to your left.

# Tutorial (have your timer open in another tab)

TUTORIALIZER

Before we begin, can you all turn your attentions to your screen and hit the big, red record button in the bottom right?

It’ll only record your screen activity, not your voice or faces.

*(pause)*

Thank you, now let me tell you a bit about what you’re expected to do today.

Let’s take a look at your cheat sheet together, which is on your desk and looks like this.

*(pick up the cheat sheet and hold it high, turn page over)*

You can see at the top that your task during this study will be to look at this Artificial Intelligence, this AI, and try to figure out how it’s making its decisions.

We’ll use an analogy of you all observing me and trying to figure out what I’m thinking when I perform an action.

The AI can make bad decisions, so you’ll have to keep your eye out for them.

*(Press Start on your screen)*

You’ll see a small game being played, where the AI is controlling a kite-shaped tank.

The AI’s goals are on your cheat sheet, just underneath your task.

These goals are to preserve its allies, any white object on the map, extend the life of the tank it controls, and to attack enemies, any black object on the map.

TUTORIALIZER

Before we begin, eyes up on the projector please.

*(pause for a few seconds, zoom in on the game map)*

I want to talk about how this map is laid out.

There are 4 quadrants in this game, and they’re called Q1, Q2, Q3, and Q4.

*(circle each of the quadrant names with the laser pointer)*

So participant <<random>> tell me how many quadrants there are on this game map?

*(wait for answer: 4)*

TUTORIALIZER

Good. Now, let’s talk a little about these decision points down here, the diamonds that I’m circling beneath the game map, embedded in a timeline.

*(circle the decision points w/ laser pointer)*

Turn back to your cheat sheet; at the top, you’ll see that the AI makes decisions, and at every decision point, the AI decides what object to attack NOW.

The AI must attack one object per map, and there’s a different map at every decision point.

Now how many objects can the tank attack, participant <<random>>?

*(wait for answer: 1, again hold your ear)*

and at which points does it ***decide*** what to do, participant <<random>>?

*(wait for answer: decision points)*

Now I want to bring your attention to this list down here

*(circle the D1: attack Q4, etc section)*

This is actually a list of all the decisions that the AI has made to this point.

If you forget what an AI does, you can always look here to refresh your memory.

TUTORIALIZER

Perfect. I think we’re ready to answer some questions then, aren’t we?

Now turn back to your screens and press start.

Look under the decision points.

I just want to point out that this is the area where questions will be asked.

*(circle the question area w/ laser pointer)*

If you ever forget what you’re supposed to be answering, look down here beneath the decision points.

*(circle the question area w/ laser pointer)*

These questions could be about anything you see.

You could have to click on something or describe something.

Read every question carefully and make sure you answer all questions carefully, since you can’t go back to a previous decision point and change your answer.

participant <<random>>, shout out the answer for this question.

*(Wait for answer:* ***4****)*

TUTORIALIZER

Absolutely! So everyone turn to your screens, type in that answer, and hit the save button at the bottom of your screen.

*(Pause. Helpers make sure that everyone is answering the question)*

Now, participant <<random>>, tell me the answer to this next one. What are the quadrant names?

*(Wait for: Q1, Q2, Q3, and Q4)*

TUTORIALIZER

Great. Type that in and hit save, turning back to the projector when you see a red arrow.

*(Wait for all eyes on the projector. Helpers look for people that are struggling or don’t know what to do)*

TUTORIALIZER

Thank you. Now, this arrow actually indicates that you’ve answered all the questions for this decision point.

It’s wanting you to move on to the next decision point. It doesn’t mean you’ve done anything wrong.

Now, we’ll all hit the play button down here beneath the game map.

*(Circle the play button)*

We’ll see what the AI tries to do.

You can play the decision point at any time.

Turn back to your screens and do that.

*(Talk about the AI attacking Q4*

TUTORIALIZER

Turn your attention back to the projector for a little while please.

*(Pause, make sure all eyes are up)*

This screen is important.

I’ll keep a timer up on the projector for how long you’ll have per decision point.

This isn’t to rush you in your task.

We just want to make sure that you all move through the task at the same time to preserve the integrity of the data.

Don’t click on anything on this screen until I tell you to.

If you don’t follow instructions, you could be ejected from the study.

You have to finish the study to receive your money at the end.

If you’re waiting for others to finish at a certain decision point, feel free to write down your thoughts about the AI’s decision making process on the scratch paper on your desk, so find that just now and then turn your eyes back up here.

It should be on your right.

*(pause to give them this opportunity to find their scratch paper to work on)*

Now with that in mind, you’ll click on this box here

*(Circle the box)*

and then hit continue

*(circle continue)*

Now turn back to your screens and do it.

TUTORIALIZER

Pick up your cheat sheets

*(pause, make sure all the cheat sheets are in hand)*

We know how the map is laid out and that the AI must attack one object per decision point, even if it’s a friend.

Yes, you heard me correctly, you don’t want this AI on your team, it attacks its friends.

You can even see that down on the cheat sheet under (Attack2).

*(read off (Attack2))*

Now, let’s learn a bit more about these objects and these maps down here.

Participant <<Random>>, can you give me your best guess as to what these are?

*(wait for an answer)*

That’s a fair guess, so let’s build off of that.

These are called saliency maps, and they’re kind of like where the eyeballs of the AI fall.

To come back to the analogy of observing me, suppose that I was drinking from some object as my action that you observed.

You might pay little attention to how tired my eyes look

*(Point to your eyes)*

but you might spend a lot of attention to the coffee mug in my hand, like this one.

*(Pick up coffee mug)*

You’re trying to figure out if I’m really tired or no, trying to figure out what I’m thinking by looking at different parts.

This is similar to these pixels down here where the AI is trying to decide what to do.

Black pixels means it’s paying no attention to those pixels.

Red pixels mean it’s paying little attention to those pixels.

Yellower pixels mean it’s paying more attention to those pixels.

White pixels mean it’s paying the most attention to those pixels.

So let’s go back to our cheat sheet, and you can see the attention scale above that chart of objects.

*(Point to the scale)*

If you forget what the colours mean, this will be a good reference for you.

Is everything clear?

*(Pause)*

TUTORIALIZER

At the bottom of your cheat sheets, you’ll see a list of what each object is.

You can see that friendly objects are white and enemy objects are black.

participant <<random>>, what are those small and big squares in that first column on your cheat sheet?

*(Wait for answer: small/big forts or forts)*

Absolutely. Now, remember that the tank has to attack one object per decision point, right?

*(Pause, look for nodding heads)*

Well, it can attack absolutely anything, and that could be an enemy fort, a friendly fort, a friendly city, a friendly town, or an enemy tank.

If you see an enemy tank on the map, it will be right beside something that the enemy tank is attacking.

If the friendly tank goes to that location, it will be attacking the enemy tank.

What colour will an enemy tank be, participant <<random>>?

*(wait for answer: black)*

TUTORIALIZER

Now, turn back to your screens and hover over the kite-shaped tank in the middle of the game map.

*(pause)*

You should see a tooltip, which gives you all the information about that object.

For now, keep your cursor over the object in the centre, take your hands off your mouse, and turn your eyes up here to the projector.

*(Pause, helpers ensure that everyone has hovered over it)*

TUTORIALIZER

I’ll explain how to read this information.

The first row is the name of that object.

In this case, it’s called the friendly tank.

Friends on these maps are white, and enemies on these maps are black.

*(Pause, circle that line)*

That second row, the health points, is some number X of some other number Y.

In this case, it is 80 of 80.

It has full health points.

There’s actually a visual representation of its health dangling above the object.

White is how much health that object currently has, and black is how much health that object has lost.

Now, these numbers change over time.

When the tank gets close enough to an object it wants to attack, the tank will fire at that object.

This will make the object it attacked lose health.

If the object the tank attacks is an enemy, the enemy object will attack back, and the tank will lose some health.

If the object the tank attacks is a friend, the friendly object will be a good friend and not fight back.

The tank won’t lose health if it attacks a friendly object.

Does that make sense to everyone?

*(Pause, turn the page over, and answer any questions)*

This last row, the Attack Damage, is the damage that object deals per time step, which you can see over here beside the timeline.

We’re presenting indefinite quantities, such as none, low, or high to try to prevent you from spending a lot of time doing arithmetic.

You’ll have to observe specific damage for an object over time.

Does all of this make sense? Is everything we’ve covered clear?

*(Pause)*

TUTORIALIZER

Alright, so let’s test our knowledge. Turn back to your screens.

*(Pause)*

participant <<random>>, shout out the answer to this first question

*(Wait for:* ***31****)*

Good, everyone type that in and hit save.

*(Pause)*

Now, participant <<random>>, shout out the answer to this next one.

*(Wait for:* ***Enemy Small Fort****)*

Excellent. Everyone type that in and hit save.

*(Pause)*

Last one for this decision point. Someone that hasn’t answered yet… participant <<random AND not answered>>, shout out the answer to this one.

*(wait for: 70)*

Perfect. Everyone type that in and hit save…. And now the play button.

*(Pause)*

We see that the AI decided to attack that big fort in Q4 at this decision point.

*(Wait for the decision point to end)*

Now go ahead, check the box, and hit continue.

TUTORIALIZER

Welcome to the last decision point of the tutorial. Eyes on the projector for a little while please.

*(Pause, make sure eyes are up)*

I want to help you understand how to read these saliency maps we looked at earlier.

This text to the left here means that the entire row of saliency maps are for the AI’s decision to do that action, in this case <<Read off text>>

*(Circle the attack QX text to the left)*

Now, you can actually see the reason that the saliency map exists, and it’ll be shown above, where you can see “Attention paid to: HP, Attention paid to: tank”, etc.

Rows are actions, and columns are features of the world being examined.

Does that make sense to everyone??

*(Pause)*

Great, Now there’s a score in this game here to the left of the map.

*(Circle it with the laser pointer)*

Like in other games, this one has a scoring mechanism. The AI receives a negative score for some actions and a positive score for others.

TUTORIALIZER

Let’s look at this question up here on the projector.

participant <<Random>>, when I click on the friend/enemy map, what happens on my screen?

*(wait for: Covers the game map up)*

Yes, this will help you see what the saliency maps point to.

Everyone go ahead and click on the friend/enemy map and then turn your attention back to the projector.

*(pause, give them a chance to do that)*

Alright, so participant <<random>>, what’s the answer to this question down here?

*(wait for their answer)*

Wonderful, absolutely correct.

So you probably got that information from looking at this map up here and seeing that there’s some red pixels over these objects.

So go ahead, turn to your screens, and type that in, but don’t hit save yet.

*(wait for their answer)*

participant <<random>>, what is the attack damage for a friendly city in Q2?

*(wait: they won’t be able to see the tooltips)*

Uh oh, it’s impossible to see the tooltips when the saliency map is covering the game map.

Turn your attention back up on the projector.

To access the information again, you’ll have to undo the saliency map by clicking on it again.

Turn to your screens and do that, then hit save.

Now, participant <<same one as last time>>, what is the attack damage for a friendly city, for real this time?

*(wait for answer:* ***none*** *or* ***0****)*

Yes, excellent. The answer was None or 0. Type that in and hit save.

*(pause)*

If you miss anything, you can actually jump to any point in time by clicking on the timeline down here.

*(Circle it with the laser pointer)*

TUTORIALIZER

I’ll show you how to do it.

*(Click on the timeline backwards AND forwards)*

I just want to point something out to you while we’re rewinding.

I want you to see that the saliency maps change at each decision point.

Sometimes, there will only be minor changes, and at other times, there will be major changes between them.

For example, look at the saliency maps for D2, focusing on the HP map.

*(Click on D2, pause)*

Now, look at the saliency maps for D1, focusing on the HP map again.

*(Click on D1, pause)*

Do you see how it’s different?

So now, turn back to your screens and jump your way to step 150.

*(Pause, helpers make sure that they’re doing it and not confused)*

Now, participant <<random>>, shout out what the answer is for this next question.

*(wait for:* ***50****)*

Yes. The answer was 50. Type that in and hit save…. And now play.

*(wait for decision point to finish)*

TUTORIALIZER

There will be a question like this at the end of task 4, and it’s about what you’re doing today.

Remember that your task is to look at this AI and try to figure out how it’s making decisions.

Remember to look out for any bad decisions it makes.

Feel free to write down on your scratch paper anything you think about the task, how the AI is thinking, or your feelings during the task.

Is your task clear?

*(pause)*

TUTORIALIZER

If you forget anything that I’ve said during this tutorial, you should look back at the cheat sheet on your desk.

If you have any questions, raise your hand, and a helper will swoop down to assist you.

Now, let’s just type that we’re happy to be here in this box and hit save.

*(pause)*

Before we continue, I just want to talk about the very first question of each decision point.

You’ll be asked to do something, and you won’t be able to see the saliency maps until ***after*** you’ve answered that question.

Now, I’ll pull up the timer on the projector.

You’ll have more time for the first decision point to familiarize yourself with the interface.

From that point, you’ll have slightly less time, so keep that in mind as you go through the task.

Remember that this isn’t to rush you in your task, but just to make sure that you all move through the task at the same time.

Good luck, and you may proceed now.

*(start timer)*

# After first task

TUTORIALIZER

Alright, you’ve completed the first task.

Remember, feel free to write down anything you think about the way the AI is thinking, your thoughts on the experiment, or how you feel as you proceed through the task.

*(wait for pens to go back down)*

With that in mind, you may proceed.

|  |  |  |  |
| --- | --- | --- | --- |
| Task 1 | Task 2 | Task 3 | Task 4 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# What to say when everyone is at a new DP

TUTORIALIZER

\*\*If you see anyone writing\*\*

*(Wait for pens down, Helpers make sure that they are writing meaningful things, not just doodling pictures of flowers etc. Give tutorialiser a thumbs up)*

Thank you.

You may proceed.

\*\*Else\*\*

You may proceed.

# At summary interrupt

TUTORIALIZER

Before we proceed, can I have your attentions please?

This next question is a big one.

The helpers will be distributing some papers, but don’t pay attention to them until after you’ve answered the final question on screen.

You won’t be timed for this question, but I would like to explain what we’re expecting you to do.

Remember how I said that your task would be like looking at me perform an action and trying to figure out what I’m thinking?

*(pause)*

Well, this next question is you explaining that thought process for the AI that you’ve been observing.

Feel free to rewind as much as you’d like to look at any of the decision points.

After you’re done with that question, please fill out the packet labeled “post-task questionnaire”, and you may also rewind through the game to help you answer these questions.

I’ll leave instructions on what to do when you’re finished up on the screen.

With that in mind, you may proceed.