Smart Applications 2020/21

Product evaluation report

Team members: L.Quarantiello, G.Tenucci, S.D.Motoc

15.01.2021 / Beta version

Product characterization

The product that we have developed is an AI able to play the social deductive game named Among AIs.
Our aim was to build an agent that could understand the world around it and act autonomously, with the goal of scoring as many points as possible. In order to do so it has to:

- move around the map following an "intelligent" path to reach its target, while trying to avoid dangerous zones, which are the ones where its enemies are located
- try to predict the enemy movements and shoot them as soon as possible
- try to reach the enemy flag and catch it
- use the chat in a friendly and human-like way
- try to play as an impostor without giving clues to its teammates
- try to understand which player is human and which one is an AI agent

We have tried to build the agent so that it acts in the most human-like way, to show to the users an as natural as possible behavior, in order to avoid being detected as an AI and to offer an enjoyable experience.
To attract a wider range of users, we have also implemented several difficulty levels; we do that by just changing the timer of the agent, so that at a lower difficulty the AI will perform the same actions, but slower.

Preparation

In order to prepare the end to end evaluation we examined the user stories and, by analyzing them, we found out that the main needs our users have involve an agent that is able to:

1. *interact with the world and play the game*
2. *be competitive*
3. *have different levels of ability*
4. *understand chat messages and commands*

From the first point, we derive that the agent should be able to move in the world, shoot and be aware of the objects around itself, *i.e.* flags, walls, barriers, energy boosts and traps.

The second point, instead, is about the way the agent performs all these actions; to be competitive it has to be “*intelligent*”, and so it should operate in an efficient way, select the best operation to do at each time, while getting in the way of its opponents.

The third one is about giving the user the possibility of selecting the strength of the AI agents, in order to have an easier or a more challenging game, according to his skills and preferences.

Lastly, the agent should understand human commands from the chat and act accordingly, to let the players organize themselves. It should also be able to keep a simple conversation about what happened during the match.

Based on the analysis above, we have extracted four measures of user satisfaction. The first one measures the general entertainment of the user playing with our agents, while the other ones are more focused on specific aspects of the AI. In particular we want to evaluate the “*strength*” of the agent, its adaptability to different skill levels and its ability to simulate human behavior. The details are the following:

1. *Have an enjoyable match against an AI player*
    a. ***magnitude**: enjoyability*
    b. ***metric**: x out of 5*
    c. ***procedure**: answer from a survey, on questions about fun, interaction with the allies, immersiveness and duration of the matches*
2. *Competitiveness*
    a. ***magnitude**: agent ability*
    b. ***metric**: x out of 5*
    c. ***procedure**: answer from a survey, on questions about strength and smartness of the agent*
3. *Adequacy of the difficulty levels*
    a. ***magnitude**: difficulty adequacy*
    b. ***metric**: x out of 4*
    c. ***procedure**: answer from a survey, on questions about the adequacy of the difficulty levels*
4. *Turing test*
    a. ***magnitude**: humanity*
    b. ***metric**: x out of 5*
    c. ***procedure**: answers from a survey, on questions about the level of human-likeness of the agent*

5. *Degree of fidelity*
    a. ***magnitude**: fidelity*
    b. ***metric**: x out of 4*
    c. ***procedure**: answer from a survey, on questions about the probability of choosing to play again and recommending the game to others*

Execution

Starting from what we defined in the Preparation phase, we designed a survey, then submitted it to the users we asked to play the game, and lastly analysed the results.

Measurement campaign

The measurement campaign was executed over a 10-days period. Given the circumstances which didn't allow us to organize a focus group to collect the data we needed, we have asked our friends and family members to play some matches against our agent and then to fill in a Google Form questionnaire.
The details about the campaign are the following:

- ***timeframe**: 03.01.2021 / 12.01.2021*
- ***reference population**: friends and family*
- ***specific procedure**:*
    - *after making each individual of the reference population play a number of games against our AI, we made them fill a questionnaire to evaluate their satisfaction*
    - *to quantify the magnitudes, for each of the ones we defined before, we took the questions related to it and applied the mean to the answers*

Collection

The metrics were collected using a Google Form questionnaire available at this link. We have formulated the survey in Italian, to make it understandable for the entire reference population. However, in the following section, we report the questions and the results in English.

In the first part of the survey, we performed an ethnographic study, by collecting some basic information, such as age, occupation and average time spent playing video games. These data will be useful to contextualize the results we will gather, in order to perform a broader demographic analysis.

The second section is instead more focused on the aspects of the game and the agents' behavior. We referred to the previously defined magnitudes to ask targeted questions, so that we could obtain a metric as accurate as possible.

In particular, for the first magnitude, *enjoyability,* we asked to evaluate some aspects of the matches, such as:

- **Entertainment**: how fun the game was
- **Interaction with the allies**: how natural the interaction with the allies was
- **Immersiveness**: measures the coherence of the actions performed by the other players
- **Match duration**: adequacy of the match duration

All these questions were rated between 1 and 5 and the final value of the *enjoyability* magnitude was obtained with the average of the results.

For the second magnitude, *agent ability,* we asked to give a rating to the quality of the agents, from both teams:

- **Strength**: how powerful the agents are
- **Smartness**: the evaluation of the strategies used by the agents

All these questions were rated between 1 and 5 and the final value of the *agent ability* magnitude was obtained with the average of the results.

For the third magnitude, *difficulty adequacy,* we requested an opinion on the difficulty levels:

- **Most used difficulty level**: We first asked the user to select its most used difficulty level, to understand which level he is rating. We provided four different levels, from *easy* to *extreme*
- **Adequacy**: Then, we inserted a question to evaluate the adequacy of such difficulty level with respect to the player's skills. Here we have 4 different answers, from *Completely adequate* to *Completely inadequate*

The *difficulty adequacy* magnitude has values between 1 and 4 and it is obtained from the second question, while the first one is used to distinguish among the several levels.

For the fourth magnitude, *humanity,* we asked the players to give an insight on how good the agents were at simulating the human behaviour:

- **Turing test**: we asked whether the user considered its allies and opponents to be AIs or human players. We gave 5 possible answers, ranging from *All humans* to *All AIs*
- **Cooperation**: we asked the user to evaluate how good were his teammates at cooperating with each other

All these questions were rated between 1 and 5 and the final value of the *humanity* magnitude was obtained with the average of the results of Turing tests and cooperation.

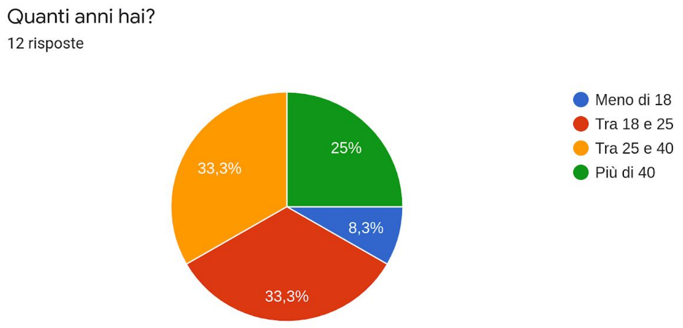
For what concerns the last magnitude, *fidelity*, we gather some information about the probability of the player returning to the game and his intention to recommend the game to his friends. For that, we asked:

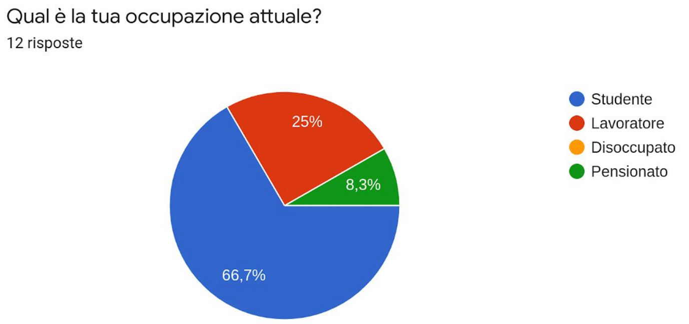
- **Playing again**: how probable the player will return to the game. We provided 4 alternatives, from *absolutely yes* to *absolutely no*
- **Recommending the game**: we asked the user to evaluate the likelihood of he/she recommending the game to a friend, ranging from *absolutely yes* to *absolutely no*

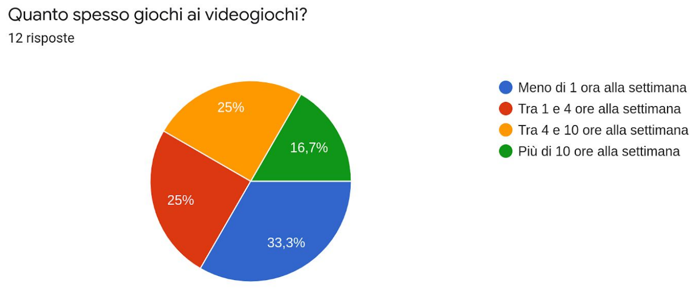
All these questions were rated between 1 and 4 and the final value of the *fidelity* magnitude was obtained with the average of the results of Playing again and Recommending the game.

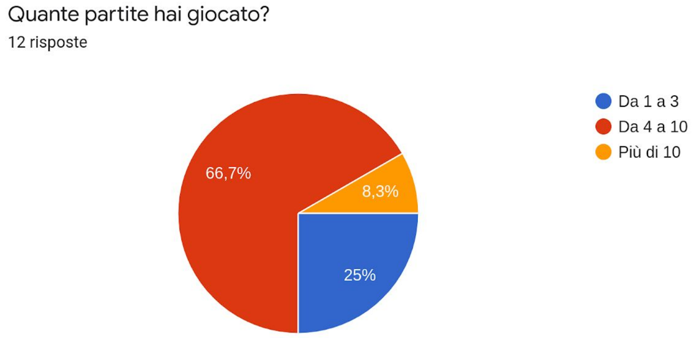
Survey results

The survey was provided to some of our friends and family members, for a total of 12 people. Every one of them had to play a few matches before compiling the survey. These are the results of the survey, as reported from the Google Form answers page.



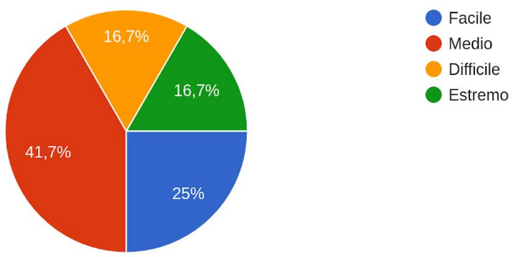






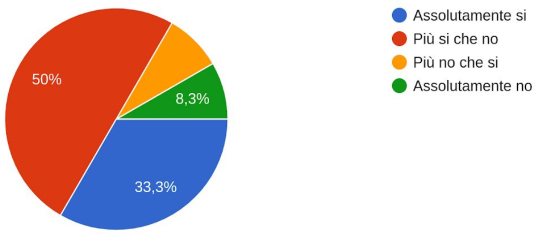
A quale livello di difficoltà hai giocato più spesso?

12 risposte

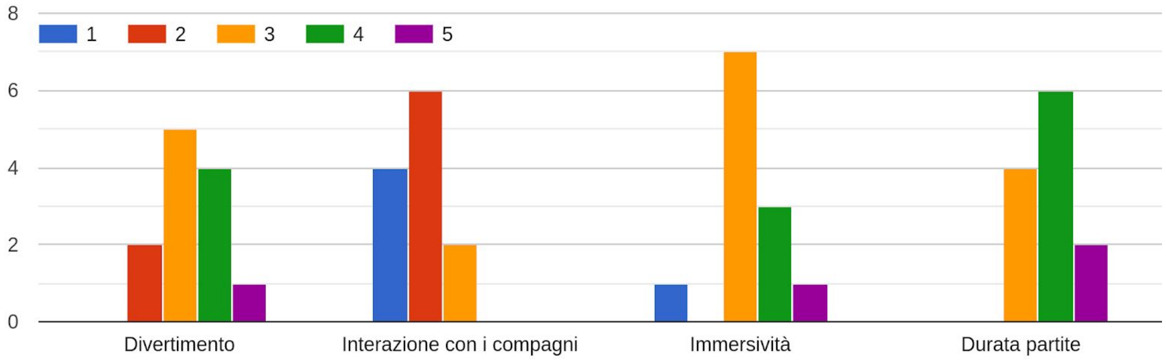


Hai trovato un livello di difficoltà adeguato alle tue capacità?

12 risposte

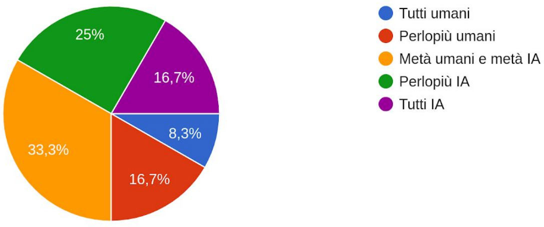


Come valuti i seguenti aspetti delle partite?



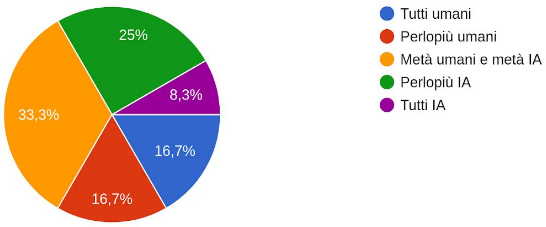
Secondo te, i tuoi compagni erano:

12 risposte

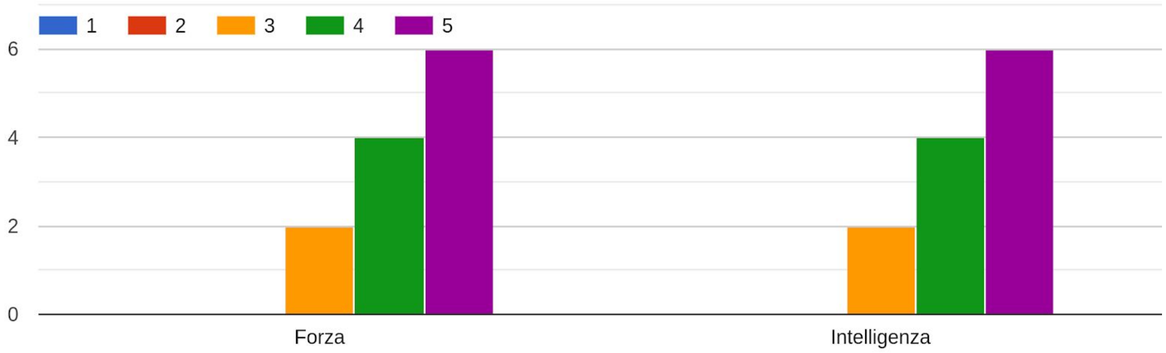


Secondo te, i tuoi avversari erano:

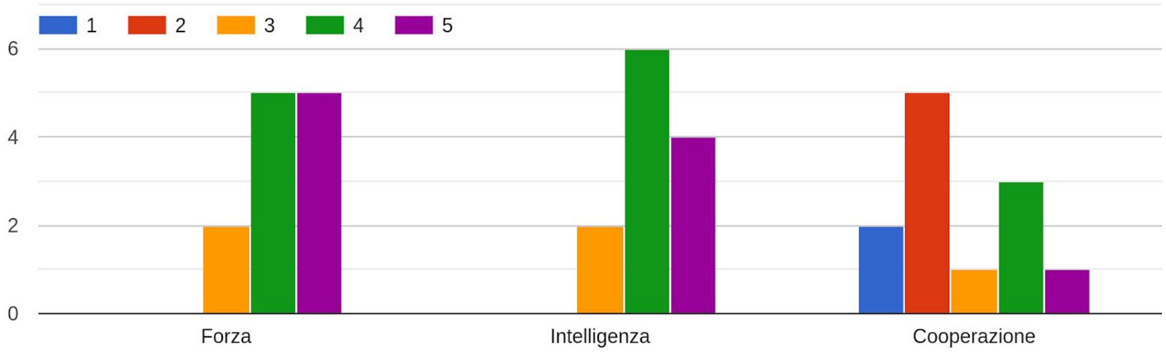
12 risposte



Come valuti i tuoi avversari?

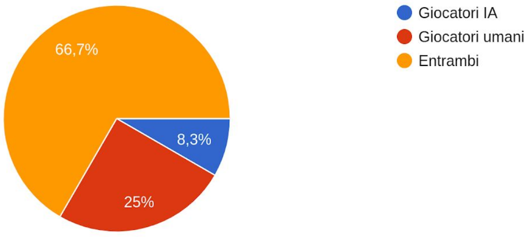


Come valuti i tuoi compagni?



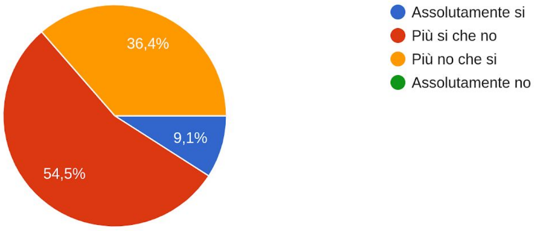
Giocheresti ancora con una IA o preferiresti giocatori umani?

12 risposte



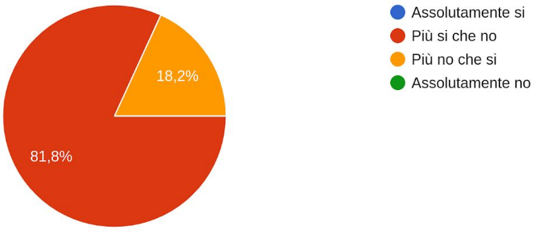
Rigiocheresti a questo gioco?

11 risposte



Raccomanderesti a qualcuno questo gioco?

11 risposte



Analysis

As we can see from the survey results, the reference population was, as expected, biased, given the fact that we could submit the survey only to friends and family members.
Most users were students, but no pattern has emerged in their playing habits as their values were spread among all choices.

Averaging the numeric results, we have:

- **Fun**: 3.33 / 5
- **Interaction with allies**: 1.83 / 5
- **Immersiveness**: 3.25 / 5
- **Match Duration**: 4 / 5
- **Opponents Strength**: 4 / 5
- **Opponents Smartness**: 4.33 / 5
- **Allies Strength**: 3.5 / 5
- **Allies Smartness**: 3.5 / 5
- **Allies Cooperation**: 2.5 / 5

From what we can see, the average user is very happy with the match duration, and rates highly the strength of our AI agents. On the other hand, he is not satisfied with the interaction with our agents, and their cooperation ratings are below average.
Oddly enough, according to the majority of the players, the opponents seem to be stronger than the allies most of the times.

Based on these results, the computed magnitudes have the following values:

- **Enjoyability** = 3.1 / 5
- **Agent Ability** = 3.83 / 5
- **Difficulty Adequacy**: 3.08 / 4
  - Easy: 2 / 4
  - Medium: 3.6 / 4
  - Hard: 3.5 / 4
  - Extreme: 3 / 4
- **Humanity** = 2.94 / 5
- **Fidelity** = 2.79 / 5

From these results we can see that our reference population has a quite good overall opinion of the game (3.1 / 5) and they consider our agent a pretty strong and smart player (3.83 / 5).
In some cases, its behavior is also considered to be human-like. In fact, the user that approaches the game for the first time is, on average, unable to distinguish between human and AI players (~3 / 5).

For what concerns the difficulty levels, we can see that on average they are considered adequate enough (~3 / 4). In particular, the middle levels, *Medium* and *Hard*, have by far the highest rating (> 3.5 / 4), while the *Easy* one seems to be quite disappointing.
Lastly, the *Fidelity* is slightly below average, so most of the times users will sadly abandon the game.

Two users also decided to submit suggestions. Their messages (translated from Italian to English) were:

- *"Some matches were not balanced. Also, I'd like to play this game on my phone"*
- *"I can't read the chat very well"*

Actions

As we can see from the results, The *Easy* difficulty is probably not easy enough for inexperienced users. Because of that, we would have to add at least another difficulty level below that, either by increasing the timer even more, or by finding other ways to cripple our AI. For example, we could decide not to take safe paths, exposing the agents more to enemy fire, or we could take largely inefficient paths to reach the opponent flag.

Since the NLP module is incomplete, our agent is not very believable with prolonged chat exchanges, so that explains the bad ratings to the interaction with our AI.
In general, though, it's hard to look very believable in chat, as users have little time to write on it, so if an agent is very active with chat messages it will probably look "*suspicious*".

We also received low scores in the cooperation section. This means we would have to find better and smarter ways of performing actions, not only according to our opponents' positions and proximity to us, but also according to what our allies are doing. For example, if many allies are storming the opponent flag, it is probably better to stay behind and defend our flag instead of joining them in the attack.

Conclusions

Due to the current situation, we could not perform a proper end-to-end evaluation campaign, but we had to rely on a small group of friends and relatives, whose votes were probably biased. Furthermore, we could only submit them a survey, while in a normal situation we would have made a focus group, to gather more information and opinion about specific aspects of the product.
In the end, this evaluation campaign was quite interesting, since we could see some reactions from outsiders on our work for the first time. Overall, the opinions about the behavior of our agent were positive, which was pretty satisfying for us.