Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created as part of a human subject study testing the impact of AI Directors in the video game test bed FarmQuest [link]. This data is a set of telemetry of all the players that played the game during that time. This dataset was created to fill the gap in game-playing datasets. There are many genres of games, and in order to more fully understand the various aspects of these genres, datasets on different genres of games should be made available. This dataset specifically targets the genre of “cozy simulation” games.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This data set was created by [RESEARCHERS] with [AFFILIATIONS].

3. Who funded the creation of the dataset?

Funding was provided by:

4. Any other comments?

None

Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance in the dataset is an anonymized player and the associated playthrough data. There is survey data consisting of demographic questions and survey questions about the player experience consisting of Likert Questions on a scale of 1-5. Additionally, players played through the game twice in a row, so there are two sessions worth of player data for each player.

2. How many instances are there in total (of each type, if appropriate)?

There are 33 players, so there are 33 three instances. Each player has all of the associated survey and telemetry data.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset is a sampling of instances. This dataset does not contain all possible instances. Specifically, there are biases as can be seen in the demographic data of this dataset. The dataset is mostly men, with only a few other genders present. The dataset is also mostly people

who consider themselves to be gamers. These biases could affect the playstyle and answers of the survey questions.

4. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each data instance consists of several pieces fo data. The first piece of data is the unique ID of the player. The second piece of data is the results from the demographic questions. The third piece of data is the telemetry from the tutorial. The fourth piece of data is the telemetry results from the first playthrough of the game. The fifth piece of data is survey results directly commenting on their first playthrough. The sixth piece of data is the telemetry from the second session playthrough. The seventh piece of data is the survey results directly commending on the second playthrough. The eighth and last piece of data is the results from a comparison survey between the two play sessions.

5. Is there a label or target associated with each instance? If so, please provide a description.

The data is stored as a JSON. The first field in the JSON is the unique ID. The demographic data data can be access from “demographic_data”. The first short survey can be accessed from “short_survey_1_data”. The second short survey can be access from “short_survey_2_data”.. The comparison survey data can be accessed from “comparison_data”. All of the telemetry data is aggregated together, and can be access with “telemetry_data”. Within the telemetry data, there is an “Event:SessionStart” that can be accessed to determine which session the player is in.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

no

7. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are two main ways to split the data. The first is splitting by session, where the data from the first play through session can be compared to the data from the second playtrhough. The Second way to split the data is by AI Director, as that was the condition that was changed during the human subject study. Thus, there could be differences in the responses or telemetry based on AI Director.

8. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The timestamp data on the telemetry is unusable, as it all recorded the same timestamp for each player. This was a problem with the way the information was recorded and was not caught until after the human subject study was finished. Instead, to track time played in game please use the in game day telemetry point. Each in game day is 1 minute, so you can effectively measure the amount of time spent in game to the floor of the nearest minute.

9. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

This dataset is currently self-contained.

10. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.

No

11. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Some of the data includes free response questions where individuals could type in their thoughts. Some of these responses could be considered moderately inappropriate due to strong language, but most of the responses are appropriate.

12. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

There is some personal information included in the demographic information, but there is an extremely low possibility of identification because only the gender and reported age range are in the dataset.

13. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

There is information on reported gender and reported age range.

14. Any other comments?

No

Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

There are two types of data in this dataset. First, there is survey data, where the participants directly answered Likert and free response questions. There is also observable data, where the telemetry of the actions that the player took in game were recorded when a player took each action, or an event happened in the game.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The survey data was collected through HTML and php. The survey was created in custom HTML, and custom php was written to store the data and encrypt it.

The telemetry data was collected through unity and php. The game was built in Unity, and all data was posted to an encrypted file using the unity www api calling on a remote server. The encryption happened when the post was received by the server using custom php code.

There was also a custom script written to decrypt the data, and save as a json file.

3. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A graduate student was paid on standard wages to create the data collection pipeline. A graduate student and undergraduate students were involved in the creation of the game in unity, and both were paid standard university rates for their time.

The participants of the study were not compensated for the data collection. The participants were gathered from slack posts in university and game development companies, posts on the university class forums, and posts on university discord.

4. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected over a time period of three months in 2024. The data was created when the participants accessed the survey, so the data collection window matches the data creation window.

5. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The study was reviewed by the ethics review board at [UNIVERISTY] and was approved as [PRO00000]. The review process involved sending all relevant documentation for review, as well as providing justifications for the data that was collected, outlining of potential benefits or harms, and justifications for the population of the study.

6. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was directly collected from individuals.

7. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The participants signed a consent form at the beginning of data collection. The link to the consent form will be provided after the double blind review, as the consent form contains identifying information

8. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

The participants could remove any partial data. The participants only needed to stop the study and all partial data is removed from this dataset. Once the participants fully completed the study, there was no way to remove the data. Participants were notified of this on the consent form.

9. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including thevoutcomes, as well as a link or other access point to any supporting documentation

No

10. Any other comments?

No other comments

Preprocessing/Cleaning/Labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. Instances for which an e

All partial datasets were removed from the data study.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Raw data is not accessible in this dataset, as per ethics approvals.

3. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Partial data was manually cleaned from the dataset.

4. Any other comments?

No other comments

Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has been used in paper [ANONYMOUS].

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

There is link [LINK].

3. What (other) tasks could the dataset be used for?

This dataset has potential use for data-driven player modeling. This dataset could be used for churn modeling, player clustering, and next action prediction, among other uses.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There is minimal risk for harm. The personal data is anonymized and is at a low risk for identification for individuals.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

This dataset was collected specifically on FarmQuest, a cozy farming game. The findings from this dataset may or may not generalize to other genres, or games of similar genres.

6. Any other comments?

No

Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description

This dataset is publicly available on the internet

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

This dataset is distributed through github

3. When will the dataset be distributed?

This dataset will be released in 2024.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

This dataset will be available for use under the MIT liscence

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No

7. Any other comments?

No

Maintentance

1. Who will be supporting/hosting/maintaining the dataset?

[AUTHORS] will be maintaining the dataset

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

[AUTHORS] can be contacted at [EMAIL ADDRESS]

3. Is there an erratum? If so, please provide a link or other access point.’

This is the only release for this dataset

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

This dataset will not be updated to add new instances or other corrections, as it is data collected from a specific study.

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

The participants of the study were notified that their data would be publicly available indefinitely.

6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

This is the only release for this dataset.

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

Individuals wishing to extend, augment, or contribute to the dataset may fork the github repository.

8. Any other comments?

No