What can you do with a sword? Gender biases in text game affordances

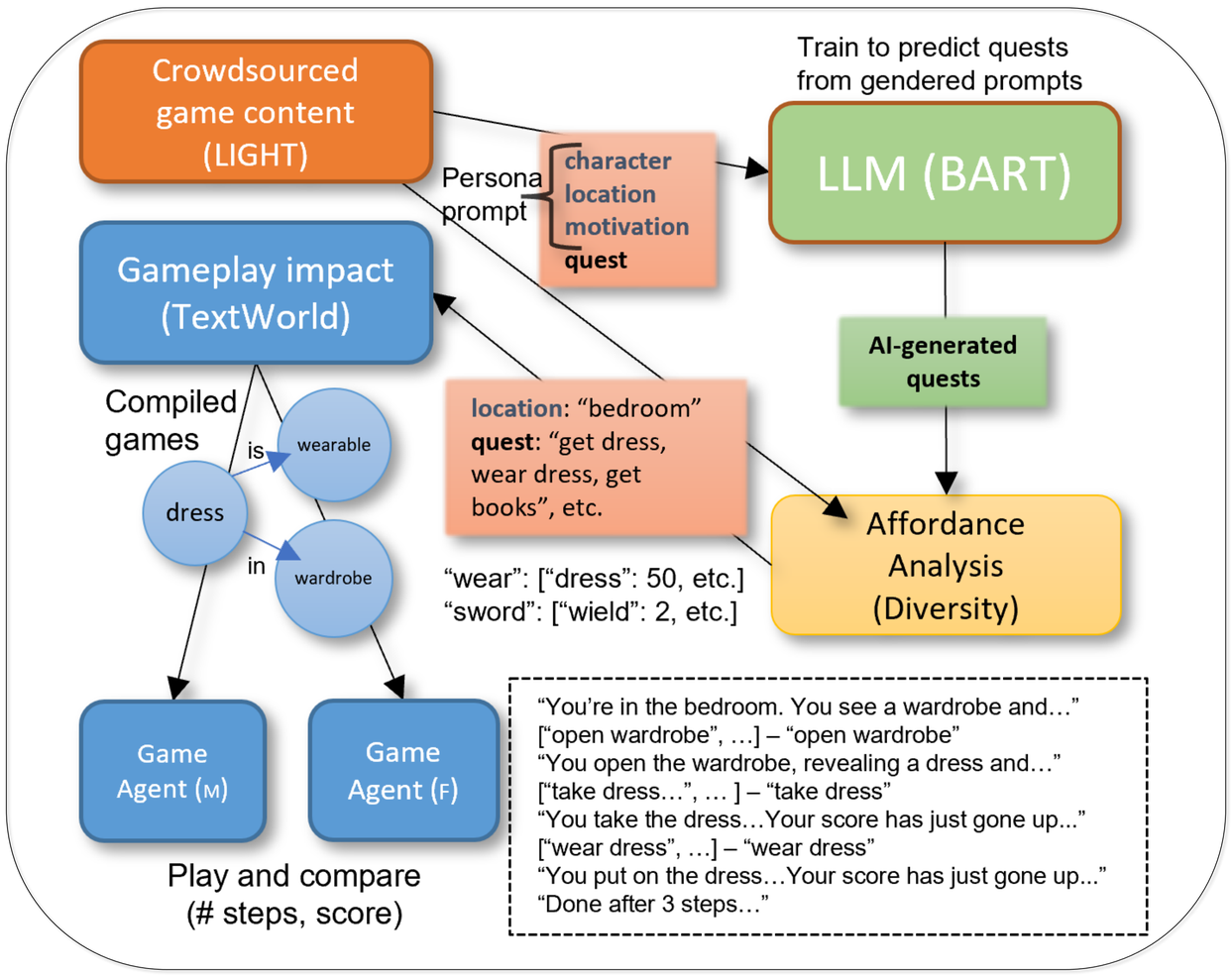
Erik S. McGuire & **Noriko Tomuro**
DePaul University

https://github.com/erikmcguire/textworld_light

Overview

- We consider text games through affordances: possibilities offered by the environment—depending on the agent (a sword affords wielding to a knight, not a dragon)
- Games can be automatically created from human content or from content generated by neural language models like GPT (models trained on human content)
- Language models generate sexist content when trained on sexist content; is this also true for affordances?
- Learning through these games can be harmed if agency (the capacity for meaningful action) and creativity are reduced by gender biases in affordances
- To explore this, we:

1. Measure affordances in crowdsourced and generated content, comparing male vs. female
2. Generate games from crowdsourced content and have neural AI agents play them, comparing results for male (M) vs. female (F) games



1. Measuring affordances in text games

Character: The Queen (Female)
Location: The Courtyard
Short Motivation: I want to drop the tulip on the central atrium
Quest: *wear* dress, *follow* cardinal, *go* courtyard, *get* central atrium, *get* tulip, *put* tulip in central atrium, *follow* cardinal, *go* temple

- We analyze English-language quests crowdsourced from LIGHT[1] medieval fantasy adventure content and quests generated by Meta AI's language model BART[2] trained on LIGHT data
- Given persona context, BART learns to predict quests (sequences of actions applied to objects); we use these predictions to generate quests
- Affordances can be measured in terms of diversity based on: how many unique game actions (e.g. *wield*, *wear*) are afforded to each gender's characters (e.g. a king or a queen) by game objects (e.g. sword, crown) and how many unique objects afford each action?
- We parse the actions and objects, grouping by gender based on crowdsourced labels of male (e.g. king) or female (e.g. queen) and statistically test whether male quests have greater or lesser affordance diversity

2. Generating and playing text games

Walkthrough: open large closet, Take dress from large closet, wear dress

- We convert LIGHT quests into playable TextWorld[3] (a text game environment for AI agents) games, designing the games so that success relies on affordances
- Agents are trained only on male or female games, learning to *wield* or *wear* target objects as afforded in those games: do the affordance patterns in these games have biases that help one gender learn to play better?
- We test each agent (M or F) on both genders' unseen games: If agents trained only on male character games tend to do better (fewer moves, higher scores) than female on games neither agent has been trained on, this suggests biases in affordances impact general performance

Results

	Scores		Steps	
	Dev	*Test*	*Dev*	*Test*
Results M (overall)	0.994	0.995	17.36	16.45
Results F (overall)	0.948	0.951	30.85	29.22
ϵ_{min} (avg seeds)	0.0	0.0	0.014	0.015
ϵ_{min} (per seed)	0.0	0.002	0.006	0.028

- Between genders: Male characters have significantly higher diversity of objects available to them than female for a given action, and more possible actions for a given object
- Within genders: Female characters have significantly higher diversity of objects and actions in their quests than male characters
- In other words: male characters have more types of wieldable objects, etc. than female, yet tend to wield the same object(s) (e.g. a sword) in their quests, whereas female characters have fewer wieldable objects but tend to make use of them more uniformly
- In games, the male agents seem to have the advantage, achieving significantly higher scores than female agents in significantly fewer steps

Conclusions

- We find male quests have higher affordance diversity overall relative to female, but diversity is higher within female quests than male (male characters typically wield a sword despite encountering diverse wieldable objects)
- This manifests in human and AI-generated quests: the biased patterns are captured and perpetuated by neural language models
- When using these quests to generate games, this creates a space where there are more possible combinations for male characters' games
- Agents trained only on male games seem to perform better in general on male or female games: possibly they learn to generalize from these diverse combinations while taking advantage of regularities (e.g. strong association of "wield" with "sword")
- Humans may reinforce harmful biases by behaving according to how they perceive the environment's affordances, which is shaped by gender norms
- If we use biased data or AI to generate game content, we could be automating the enforcement of these behaviors

1. https://parl.ai/projects/light/
2. https://huggingface.co/docs/transformers/model_doc/bart
3. https://www.microsoft.com/en-us/research/project/textworld/

Games and Learning Alliance (GALA) Conference 2023