Collaborative Storytelling with Human Actors and AI Narrators
Paper type: Event Report

Boyd Branch
University of Kent
United Kingdom
boyd@improvmedialab.com

Piotr Mirowski
Improbotics
United Kingdom
improbotics.org

Kory Mathewson
Improbotics
Canada
improbotics.org

Abstract

Large language models can be used for collaborative storytelling. In this work we report on using GPT-3 (Brown et al. 2020) to co-narrate stories. The AI system must track plot progression and character arcs while the human actors perform scenes. This event report details how a novel conversational agent was employed as creative partner with a team of professional improvisers to explore long-form spontaneous story narration in front of a live public audience. We introduced novel constraints on our language model to produce longer narrative text and tested the model in rehearsals with a team of professional improvisers. We then field tested the model with two live performances for public audiences as part of a live theatre festival in Europe. We surveyed audience members after each performance as well as performers to evaluate how well the AI performed in its role as narrator. Audiences and performers responded positively to AI narration and indicated preference for AI narration over AI characters within a scene. Performers also responded positively to AI narration and expressed enthusiasm for the creative and meaningful novel narrative directions introduced to the scenes. Our findings support improvisational theatre as a useful testbed to explore how different language models can collaborate with humans in a variety of social contexts.

Improv Theatre and AI

Improvisational theatre is increasingly being used as an environment and platform for testing and exploring the creative potential of computationally creative systems (Bruce and others 2000; Baumer and Magerko 2010; O'Neill and others 2011; Jacob 2019) and of artificial intelligence language models in particular (Martin, Harrison, and Riedl 2016; Mathewson and Mirowski 2017; Mathewson and Mirowski 2018; Cho and May 2020). Turing test inspired experiments focus on evaluating how well language models can perform natural-sounding human language (**?**). Conversely, improvisational theatre is uniquely positioned to explore the collaborative potential of language models. Collaborative storytelling includes both on-stage performance (e.g. improvised theatre) and off-stage games (e.g. table-top role-playing games, card games). Collaborative storytelling in collaboration with artificial agents has been studied previously (Perlin and Goldberg 1996; Hayes-Roth and Van Gent 1996; Riedl and Stern 2006; Magerko and others 2011), most often in the context of virtual environments where the human players interact with digital avatars, with the exception of plot generation tools like *dAIrector* (Eger and Mathewson 2018) informing live improv on stage. Recent advances in large language models enable richer text-based interaction between human and AI players (Nichols, Gao, and Gomez 2020), as illustrated by the success of online role-playing game *AI Dungeon*[1]. Our case study pulls away from the virtual world and situates AI and human collaborators together on-stage. This shared narrative is then interpreted by live human actors, expressing the full range of emotional, physical and verbal human creativity.

Improvised theatre explores how interesting narratives can emerge from establishing rules for simple social dynamics and rhetorical conventions. In contrast to scripted theatre, improv is built from spontaneity (Spolin and Sills 1963). Improvisers are trained to disengage executive cognition in order to allow their automatic responses to guide and justify a given and emerging social context (Johnstone 1979). Narratives emerge by assuming the presence of meaning. The performer only needs to accept *offers*: what is said and done on stage. There are no 'wrong' things a performer can say to invalidate the emerging narrative. For meaning to emerge for an audience however, each novel narrative statement must be followed up with some degree of agreement and justification. An improvisational scene is ultimately judged on the degree to which novel statements can be integrated back into the previous given circumstances. That process is called *justification* and is synonymous with an ongoing adaptation, by the actors–thrown out of their comfort zone–to the changing dynamics of an improvised narration. This practice is what makes improv theatre such a useful platform to explore the creative capacity of artificial intelligence. For the AI to perform 'well' it cannot simply introduce novel narrative subjects, but must also be able to adapt to the emerging given circumstances, akin to the desiderata of AI systems capable of generalising to unseen data.

Previously, (Mathewson and Mirowski 2017; Mathewson and Mirowski 2018; Cho and May 2020) explored how an

[1] https://play.aidungeon.io/

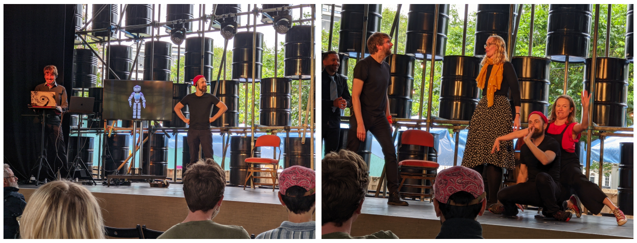

Figure 1: Left: Operator of the AI narrator and virtual avatar. Right: Example of an improvised scene. Photos: Erika Hughes

AI trained on movie or improv dialogue could generate interesting narratives as a *performer* within an improvised scene, and demonstrated that conversational agents built using recurrent neural networks or transformers, e.g., GPT-2 (Radford and others 2019), could indeed move a given narrative forward when human agents were operating to accept and justify the statements. In their setup, the AI only functioned as a character in a given scene. In our study, we examine how AI performs in the role of the narrator.

Methods

Datasets for AI improv

Before large language models, conversational agents were trained on datasets geared towards dialogue, like the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil and Lee 2011) and OpenSubtitles (Tiedemann 2009). The latter was used to develop conversational models such as (Vinyals and Le 2015) and the improv theatre-specific chatbot A.L.Ex (Mathewson and Mirowski 2017) from *HumanMachine*[2]. Later on, an improv-specific dataset of *yes-and* exchanges from improv podcasts was curated in (Cho and May 2020).

While employed in the context of improvised storytelling, our work departs from generating dialogue and focuses on storytelling from the perspective of a narrator. We hypothesized that the best datasets would come from general fiction novels as well as synopses and plot summaries. Coincidentally, Large Language Models (LLMs) are now pre-trained on comprehensive and diverse sets of corpora and are capable of memorising diverse linguistic patterns from books, novels, movie scripts, newspapers or blog posts (Radford and others 2019; Brown et al. 2020). From the perspective of thematic diversity and specificity, the need for collecting specific training data seems to have become less of an issue.

Live curation, mitigating of model bias

There are however trade-offs between the predictive power of large language models, and their embedded biases or their misalignment with desired societal values, which have been discussed in (Bender et al. 2021; Kenton et al. 2021).

Our approach to mitigating these biases and to the removal of offensive content relies on a combination of automated filters and human curation, performed in real time in the context of a live show. First, we remove sentences that contain known offensive words from a blocklist, and all generated sentences are validated using multiple filters for inflammatory, hateful or sexual content by the Perspective API[3]. Second, the human who operates the storyteller interface has agency in both how they formulate and type the context, and in what sentences produced by the AI they choose to read, with a possibility to omit or reword parts of those sentences.

Interactive live AI narration on stage

Our interface works in the following way: each time a sentence is typed by the operator, it is concatenated to the context of the scene. GPT-3 is then run 3 times on the whole context, thus generating three sets of sentences of total length up to 100 characters for each set. The operator has the choice of selecting none, one, or several of these sentences, in the order they choose[4].

An important aspect of the human-machine interaction on stage is that the actors' performance and the operation of the AI happen simultaneously, i.e. that the operator types context prompts and chooses AI-generated suggestions at the same time as scenes unfold. The human operator may then interrupt the scene, in a similar way to an improviser 'editing' the scene. This delegates to the human cast and to the operator the artistic choices of timing–a crucial element of comedy–and maintains the liveness of the performance.

Story initiation We initially experimented with a system for automated selection of initial writing prompts from the *novel-first-lines-dataset* (a crowdsourced dataset of first sentences of novels)[5]. The single-word audience suggestion would be matched with a fixed set of 11k sentences using sentence-level embeddings computed using the Universal Sentence Encoder (Cer et al. 2018) combined with approximate nearest neighbor search[6]. Early trials during improv rehearsals demonstrated that the first lines of novels were not informative enough for the actors performing much shorter scenes, and that the actors preferred to initiate the story themselves.

Avatar for the AI narrator We designed a virtual avatar that personified the AI narrator. That avatar consisted of a 3D model of a robot, inspired by Aldebaran Robotics' *Nao*, built using Cinema 4D[7] and imported into Adobe Character Animator[8] as a puppet controlled by facial expressions of the operator as they are reading the AI-generated lines. Instead of using computer-generated voice, we relied on human voice for expressive interpretation. The operator was

[2] https://humanmachine.live

[3] https://www.perspectiveapi.com/

[4] While it has been observed that GPT-3 is capable of some amount amount of meta-learning, such as recognising and generating analogies, or responding to *commands* (e.g., ”translate the following sentence from English to French”) (Brown et al. 2020), we decided to limit this work to using GPT-3 as a statistical language model and to leave, for future work, hierarchical text generation or additional prompt engineering, such as “expand on what happens next” or “let's look back at this character”.

[5] https://github.com/janelleshane/novel-first-lines-dataset

[6] https://github.com/korymath/jann

[7] https://www.maxon.net/

[8] https://www.adobe.com/products/character-animator.html

standing behind the TV screen that was projecting the avatar on the stage.

Evaluating AI in performance

We worked with a small team of professional improvisers to build and rehearse an original 50 minute performance that included a series of short AI-assisted scenes followed by a 12-minute AI narrated long-form improvisational scene. We then presented 2 performances for public audiences. We present a partial transcript from one of the AI narrated performances and discuss how well the AI was able to offer contextually relevant suggestions that advanced the plot. We also administered anonymous surveys to the performers (p1-p5) as well as 9 audience members after each performance (a1-a9). The surveys consisted of a series of open ended questions regarding how they experienced the AI on stage. Our surveys were conducted in accordance with the approved ethical standards of our public research institution.

Results

We relate a 12-minute long-form improvisation between 6 actors. GPT-3 generated altogether 455 sentences of suggestions, but only a subset was selected by the operator. The following extract is an exact transcript of only the inputs or *context* given to the GPT-3 language model. **Sentences in bold correspond to GPT-3 outputs, which are fed back again as context**. *(Notes in italic and between parentheses are scene descriptions that were not input to GPT-3).* The transcript of performer lines and interactions is not included because of the difficulty of transcribing an improvised performance where many actors may speak on top of one another. The audience gave the suggestion: “Pizza Hut”.

> At the Pizza Hut. Brian and his date lost patience. *(The operator misunderstood the relationship between the two protagonists.)*
>
> There was always a reason for them to admire each other. Brian was an expert at making pizza. Sally found her vocation, making pizza like Brian. Brian started listing all the products... Baguettes, patisserie... Sally asked Brian for help. *(The operator made a confusion in the name, as it was Sandra, not Sally.)*
>
> **The door opened and a burly man entered, followed by his wife.** *(A couple entered the pizzeria, the man spoke with a heavy voice.)*
>
> The husband and the wife entered the pizzeria. They asked for supremes, with garlic bread. **Both women had crushes on Brian.** *(The unnamed wife briefly approached Brian.)*
>
> Sally searched for pastries. **The husband and the wife asked for vodka.** *(Unused suggestion.)*
>
> They got creme patissiere... **Brian apologized.** *(Sally/Sandra was rolling pizza on the floor.)*
>
> Sally was dreaming about becoming a master patissier. **She continued to look for pastries.** *(Sally/Sandra said she was done working at Pizza Hut and wanted to resign. Scene transition, with an angry boss entering the stage.)*
>
> Brian’s boss told him he would let her go. Sally gave her notice. The boss refused. The boss was cruel. **Brian asked the boss for her resignation. The boss made a mistake.** *(A confrontation took place between Brian and the boss, the boss later started behaving apologetically.)* **Brian and Sally left the pizzeria.** *(A male actor stepped in to play the newly introduced Sally.) (Scene transition to Sandra at a restaurant owned by the burly man and his wife.)*
>
> Sandra pursued her dream of being a pastry chef. Sandra was serving the old burly couple. **The burly man was impressed. The burly man and his wife complimented Sandra.**
>
> Even though Sandra was violating safety regulations. **Sandra was getting tired. Sandra’s dream would soon come true.**
>
> They loved it! With her sweat, she impressed them. **Sandra was now a great pastry chef.** *(Scene transition to the boss joining the group.)*
>
> The boss came to apologise to Sandra. Sandra said that she remembered him. He was diminished. He was wondering if it was safe to do it on the floor... She heard about Brian. Can you come back, he asked. The boss was apologetic. **Sandra thanked the boss, who helped her. Brian and Sandra were both happy. Sandra was proud. The boss was really clear. The boss was jealous.** *(Group scene.)* **He agreed.** *(End scene.)*

As the 12-minute scene unfolded, the operator was typing a summary of it as inputs to an interface to GPT-3. For each line of context that was input by the operator, there were many alternative suggestions that could have been selected, and this transcript shows only the ones that were actually chosen and presented to the cast and to the audience. The decision to intervene in the narration and the timing and delivery of each intervention were choices made by the operator, who was simultaneously voicing and animating the virtual avatar, as well as observing the live improvised scene.

Just like in the first show (for which we do not report the transcript in this paper), the AI-assisted narrator’s interactions became more frequent as the scene was unfolding and the characters established. The motivation for this was to let the actors establish the characters and their relationships first, and to start intervening only once the cast had an initial guess of the narrative arc of the story.

Audience and performer response

We provide the following small sample of 9 audience responses as useful observations to guide discussion rather than evidence of findings that can be generalised. 7 of the respondents indicated the presence of AI itself as the most significant motivational factor in attending the events. 6 reported overall satisfaction with AI narration, 1 reported neutral satisfaction, and two reported dissatisfaction. The AI narrated scene was the most frequently cited (5/9) response to the question ‘What did you enjoy most about the show.’

All 5 performers reported satisfaction with the ability of the AI to move the story forward. 3 however also ‘slightly agreed' that the AI 'mainly introduced absurd or random information' into the scenes. We present the following quotes from performers about their experience to advance discussion about the relationship between the insertion of surprising plot points experienced as both 'random' and useful in advancing the story arc.

- *As a performer I had to physically become the character (the AI) described in the narration. This pushed me to a certain pov / voice/ physicality which I probably wouldn't have chosen i.e. a gruff, muscly patisserie store owner.* (p1)
- *(The AI) added a level of randomness and craziness different from a human brain.* (p2)
- *(The AI) really helped the plot move forward, but without being too prescriptive, and enabled me to focus on character development, relationships, emotions and object work.* (p3)
- *I did a few scenes as the protagonist where I was sad, and then (the AI) would say 'she was happy' or similar, but I loved that as I has to justify it and it was funny!* (p4)
- *Generally the narrative direction (of the AI) helped the show move forward in a good direction* (p5)

Discussion

In the above exchange between the operator's inputs and the AI suggestions, one can notice that the AI introduced two key characters (the *burly man and his wife*) who played the role of mentors for the main protagonist, *Sandra*, and enabled the resolution of the story by complimenting *Sandra*'s work. The AI's suggestions also satisfied a classical narrative arc by allowing her "dream to come true" and achieving her transformation into a "great pastry chef". This illustrates the capacity for an AI-based narrator–operating in tandem with a human curator who makes timing decisions–to generate novel and meaningful plot points.

Interestingly, as (p4) noted, the AI-provided suggestions did not consistently keep the affect or motivational stance of some characters (e.g., the boss was first cruel, then apologetic and even helping Sandra). Where this inconsistency might invalidate a progressing story when uttered from a character (and subsequently fail a Turing-test), in the mouth of a narrator it can encourage performers to maintain classical story arcs that require characters to change and adapt over time (Aristotle 350 BC). The inconsistencies of the AI-generated text were interpreted by the cast as narrated reversals of feelings, and challenged the performers (as p1 suggests) to allow themselves to change and be affected by each other in surprising but meaningful ways. In improv theatre this is described as a 'status' reversal where the 'low' status of a character at the beginning of a scene becomes 'high' status by the end (Giebel 2019). Such reversals are in practice often difficult for human improvisers to execute, as one instinctively attempts to maintain their given status or fight to maintain 'high' status. In this instance, the AI drove the plot more aggressively forward and motivated the performers to shift and adapt status to the evolving circumstance that in effect provided a more clear beginning, middle, and end to the story.

As a creative partner, rather than simply providing strange or absurd plot points to challenge the human performers to make sense out of, the AI seems to have removed some of the cognitive load for improvisers (as with p3) allowing them to concentrate on relationships. Without a narrator, improvisers must both react spontaneously in the moment, and remember to engage narrative techniques such as status changes to move the story forward. This important practice of narrative making can be understood as "a carefully argued process of removing and adding participants" (Kumar et al. 2008). The practices of theatrical improvisation and acting techniques such as Meisner actor training (**?**) explicitly ask performers "not to be in their heads", meaning not to withdraw from the live performance in order to plot or to reflect and comment on the scene, but rather to dedicate their entire attention to what is happening in front of them on the stage. We believe that one of the potential applications of computational creative systems could be to alleviate the cognitive load of performers to shift their focus from plotting to reacting.

Strikingly, the seemingly random characters introduced by the AI were often the result of human error. But even when such errors were introduced the performers playfully accepted the offers that resulted in comic relief, skilfully transforming an error into a serendipitous opportunity to 'break the fourth wall' and to connect with the audience. For instance the wrong naming of the main protagonist (as the operator mistakenly and repeatedly typed "Sally" instead of "Sandra") led the improvises to quickly introduce, then shelve, a temporary character. This tight collaboration between improvisers and AI prevented the introduction of a new character that may have otherwise been considered a 'less carefully argued' addition to the narrative, to still perform a useful function (comic relief) without disrupting the evolving story.

Conclusions

Narrative theatrical performances encapsulate human culture, social interaction, physical expression and natural human emotion. Improv is an ideal test-bed to explore questions about the human-AI collaborative creative capacity. It has been proposed as a *grand challenge* for artificial intelligence (Martin, Harrison, and Riedl 2016). We believe that AI-as-collaborator, as in this current study, uplifts artists, as opposed to challenging them.

Language models capture statistics of written corpora of human culture, and thus provide human audiences with a mirror of typical narrative tropes and biases. Thus, they highlight the need for human interpretation and curation of AI-generated content. Our two-pronged approach of automated filters followed by human operator selection of sentences, illustrates a transfer of responsibility from the language model to the (human) narrator–not unlike a typical improv show, where the human cast are responsible for the story they tell (e.g., "punching up, not down") and adapt to their audiences (e.g., family-friendly vs. late-night shows).

This work is, to the best of our knowledge, the first staging of an AI narrator co-creating improvised theatre alongside humans for a live audience. Timing and aesthetics are significant factors for the human experience of AI by audiences and cast members. The ease of use of the narrative interface for the human operator impacts how quickly they can add to the language model context or choose from its outputs. Finally, the imagined 'personality' of the AI narrator play a role in co-creation. These are important avenues for future research on human-AI co-creation.

Acknowledgments

The authors wish to acknowledge the anonymous audience participants who filled the questionnaires, and above all, to thank the actors from improv theatre troupe Improbotics[9] who collectively created the artwork that enabled this specific study, namely Harry Turnbull, Julie Flower, Marouen Mraihi, Paul Little and Sarah Davies.

References

[Aristotle 350 BC] Aristotle. 350 B.C. *Poetics*. Translated by S. H. Butcher.

[Baumer and Magerko 2010] Baumer, A., and Magerko, B. 2010. An analysis of narrative moves in improvisational theatre. In *JICIDS*, 165–175. Springer.

[Bender et al. 2021] Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Conf. on Fairness, Accountability, and Transparency*, 610–623.

[Brown et al. 2020] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[Bruce and others 2000] Bruce, A., et al. 2000. Robot improv: Using drama to create believable agents. In *IEEE ICRA*.

[Cer et al. 2018] Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[Cho and May 2020] Cho, H., and May, J. 2020. Grounding conversations with improvised dialogues. *arXiv preprint arXiv:2004.09544*.

[Danescu-Niculescu-Mizil and Lee 2011] Danescu-Niculescu-Mizil, C., and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

[Eger and Mathewson 2018] Eger, M., and Mathewson, K. W. 2018. dairector: Automatic story beat generation through knowledge synthesis. *arXiv preprint arXiv:1811.03423*.

[Giebel 2019] Giebel, J. D. 2019. Improvising tactical choices based on status or "who's driving the dramatic action bus?". *Objectives, Obstacles, and Tactics in Practice: Perspectives on Activating the Actor*.

[Hayes-Roth and Van Gent 1996] Hayes-Roth, B., and Van Gent, R. 1996. Improvisational puppets, actors, and avatars. In *Proc Comp Game Dev Conf.*

[Jacob 2019] Jacob, M. 2019. *Improvisational Artificial Intelligence for Embodied Co-creativity*. Ph.D. Dissertation, GIT.

[Johnstone 1979] Johnstone, K. 1979. *Impro: Improvisation and the theatre*. London: Faber and Faber Ltd.

[Kenton et al. 2021] Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

[Kumar et al. 2008] Kumar, D.; Ramakrishnan, N.; Helm, R. F.; and Potts, M. 2008. Algorithms for storytelling. *IEEE Transactions on Knowledge and Data Engineering* 20(6):736–751.

[Magerko and others 2011] Magerko, B., et al. 2011. Employing fuzzy concept for digital improvisational theatre. In *AIIDE*, 53–60.

[Martin, Harrison, and Riedl 2016] Martin, L. J.; Harrison, B.; and Riedl, M. O. 2016. Improvisational computational storytelling in open worlds. In *International Conference on Interactive Digital Storytelling*, 73–84. Springer.

[Mathewson and Mirowski 2017] Mathewson, K. W., and Mirowski, P. 2017. Improvised theatre alongside artificial intelligences. In *AIIDE*.

[Mathewson and Mirowski 2018] Mathewson, K. W., and Mirowski, P. 2018. Improbotics: Exploring the imitation game using machine intelligence in improvised theatre. In *AAAI AIIDE*.

[Nichols, Gao, and Gomez 2020] Nichols, E.; Gao, L.; and Gomez, R. 2020. Collaborative storytelling with large-scale neural language models. In *Motion, Interaction and Games*. 1–10.

[O'Neill and others 2011] O'Neill, B., et al. 2011. A knowledge-based framework for the collaborative improvisation of scene introductions. In *ICIDS*, 85–96. Springer.

[Perlin and Goldberg 1996] Perlin, K., and Goldberg, A. 1996. Improv: A system for scripting interactive actors in virtual worlds. In *Conf on Comp. G. & Int. Tech.* ACM.

[Radford and others 2019] Radford, A., et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

[Riedl and Stern 2006] Riedl, M. O., and Stern, A. 2006. Believable agents and intelligent story adaptation for interactive storytelling. In *Intl Conf on Tech for Int Dig St and Ent*, 1–12. Springer.

[Spolin and Sills 1963] Spolin, V., and Sills, P. 1963. *Improvisation for the theater: A handbook of teaching and directing techniques*. Northwestern University Press.

[Tiedemann 2009] Tiedemann, J. 2009. News from opus-

[9] https://improbotics.org

a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in NLP*, volume 5.

[Vinyals and Le 2015] Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.