Empirically Evaluating Creative Arc Negotiation for Improvisational Decision-making

Mikhail Jacob*
t-mijaco@microsoft.com
Microsoft Research
Cambridge, UK

Brian Magerko
magerko@gatech.edu
Georgia Institute of Technology
Atlanta, USA

ABSTRACT

Action selection from many options with few constraints is crucial for improvisation and co-creativity. Our previous work proposed creative arc negotiation to solve this problem, i.e., selecting actions to follow an author-defined 'creative arc' or trajectory over estimates of novelty, unexpectedness, and quality for potential actions. The CARNIVAL agent architecture demonstrated this approach for playing the Props game from improv theatre in the Robot Improv Circus installation. This article evaluates the creative arc negotiation experience with CARNIVAL through two crowdsourced observer studies and one improviser laboratory study. The studies focus on subjects' ability to identify creative arcs in performance and their preference for creative arc negotiation compared to a random selection baseline. Our results show empirically that observers successfully identified creative arcs in performances. Both groups also preferred creative arc negotiation in agent creativity and logical coherence, while observers enjoyed it more too.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Computing methodologies** → *Artificial intelligence*; Neural networks; • **Applied computing** → *Performing arts*.

KEYWORDS

human-AI improvisation, empirical evaluation, improvisational theatre, interactive installation

1 INTRODUCTION

Improvisation is a ubiquitous activity that humans do every day. Professional improvisers perform to crowds, regularly finding the perfect response just in time for the unfolding action on stage. While experienced improv performers make the task look easy, improvisational performance is supremely challenging for many

---

*Work done while at the Georgia Institute of Technology.

reasons [14, 28]. Research in human-computer improvisation enables us to create novel forms of artistic expression and interactive experience that combine the social creativity and engagement of improvisation with the scale and democratisation of artificial intelligence (AI). However, improvising with humans poses additional challenges for any improvisational AI agent [18].

One of the primary challenges for improvisational agents that need to improvise with humans outside of small-scale research prototypes is choosing their next action from a vast set of options in near real-time without a small, well-defined set of goals or hard constraints to optimise. This is called the *improvisational action selection* problem [9]. Failure to address this problem can result in incoherent behavior, decision paralysis, or repetitive responses [9].

Previous work [13] proposed *creative arc negotiation* as one solution for improvisational action selection inspired by perceived experiential arcs found in many creative fields. An agent using this approach is intrinsically motivated to negotiate an author-specified 'creative arc' along with a human improviser, selecting actions to best follow it during an improvised performance (see section 4). Creative arcs are authored trajectories through a 'creative space', i.e., a three-dimensional space consisting of novelty, unexpectedness, and quality estimates of potential actions. This work adapts Boden's [3] definition of creativity, paraphrased as the novelty (originality, either personally or historically), surprise (violation of expectations), and value (subjective importance or desirability) of a creative artifact. Unexpectedness and quality estimates were used as proxies for the overloaded concepts of surprise and value.

We implemented creative arc negotiation in the CARNIVAL agent architecture (section 4) and studied it within the virtual reality (VR) installation called the Robot Improv Circus [10, 13]. The installation enables participating improvisers to play an improv game with an AI agent in VR, while an audience views the performance from just outside (section 3). Participants play the Props game from improv theatre in the installation, where an improviser takes turns with a virtual agent to perform actions and dialogue with an abstract prop, pretending it to be a real or fictional object for comedic effect.

This article presents three studies that evaluate the effect of creative arc negotiation on the experience of participating improvisers and observers. We use two large-scale crowdsourced studies of observers and one smaller laboratory study with improvisers to evaluate the following questions. 1) Can observers and improvisers identify creative arcs when an agent uses them for action selection? 2) Do observers and improvisers prefer creative arc negotiation to a random selection baseline in terms of enjoyment, agent creativity, and logical coherence? We present our results and discuss their implications for human-AI improvisation.

(a) First-person view of agent miming an action with a prop.

(b) Audience view shown to spectators outside installation.

Figure 1: The Robot Improv Circus VR installation. Speech bubble reads, "I am looking with my kaleidoscope."

2 RELATED WORK

Improvisers demonstrate near real-time collaborative creativity in open-ended and poorly-defined problem domains [14, 23]. A limited number of improvisational agents exist for theatre [4, 22, 25, 26] and storytelling [20]. The Three Line Scene [25], Party Quirks [26], and Tilt Riders [4] systems are cognitive models of improvisational process, using small amounts of hand-authored expert knowledge to improvise. These systems don't explicitly address the improvisational action selection problem when reasoning over larger amounts of knowledge unlike the use of creative arc negotiation in CARNIVAL. More recent improvisational systems such as [20, 22] offer exciting solutions for improvisational action selection using larger data sets and machine learning. In addition to addressing this same problem, creative arc negotiation also enables the system to evaluate both the human's and its own actions in the moment while responding, allowing for more creative responsibility and autonomy.

Creative arc negotiation can also be considered a form of intrinsic motivation for agents to follow a given creative arc. This is similar to drives for curiosity (seeking out novel or unexpected stimuli) [24, 33] or empowerment maximisation (seeking to maximise influence over future outcomes) [32]. Similarly, evolutionary computing has demonstrated agents trained using novelty search [17] and surprise search [7], where agents are selected for achieving the most novel or unexpected outcomes instead of the highest quality outcome. Creative arc negotiation operates at a meta level to generalise these approaches beyond pure maximisation, selecting novelty, unexpectedness, and quality to follow a designer-specified creative arc instead. In principle, one could also add other motivations to the negotiation process, for example, to follow an arc of empowerment. Additionally, improvisation is not intended to generate a final product or outcome but on the experiential journey of an ephemeral performance. Therefore, motivating the agent to maximise any aspect of the experience would be counterproductive.

Creative arc negotiation has roots in experience management [2, 31] and other interactive narrative research [30] like Façade [21] and Merchant of Venice [27]. Façade sequences hand-authored story fragments or 'beats' for the player to follow an arc in dramatic tension. In principle, creative arc negotiation could also emulate

dramatic tension search by adding dramatic tension as an added dimension to the negotiation process, while also searching through novelty, unexpectedness, and quality. The Merchant of Venice research [27] describes a visual programming method for drawing dramatic arcs in order to guide a planning-based interactive narrative experience of the Merchant of Venice. This is a potentially useful idea that could be incorporated into CARNIVAL in the future to enable personalization of creative arcs between players or ease the creative arc authoring process for non-experts.

Figure 2: A participant performing in the installation.

3 THE ROBOT IMPROV CIRCUS

Improvisers usually play the Props game by taking short turns to perform quick, one-shot actions and dialogue with an abstract prop, pretending it to be a real or fictional object for comedic effect. The Props game allowed us to study short sequences of related actions between players with a common prop, without requiring them to create full narratives. A simplified version of the Props game with a limited set of abstract props and no dialogue exchange was used to study improvisational action selection in the VR installation — the Robot Improv Circus (figure 1) [13].
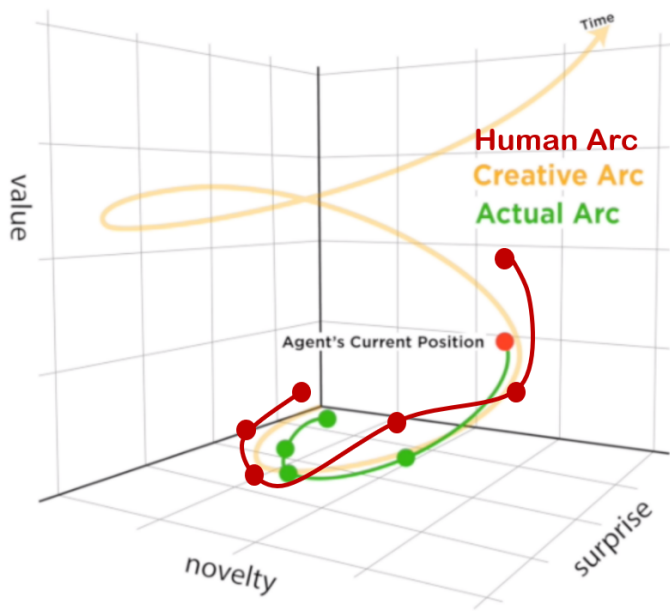
(a) The creative arc negotiation process. The agent follows a designer-given creative arc considering the human's and its own actions in creative space.
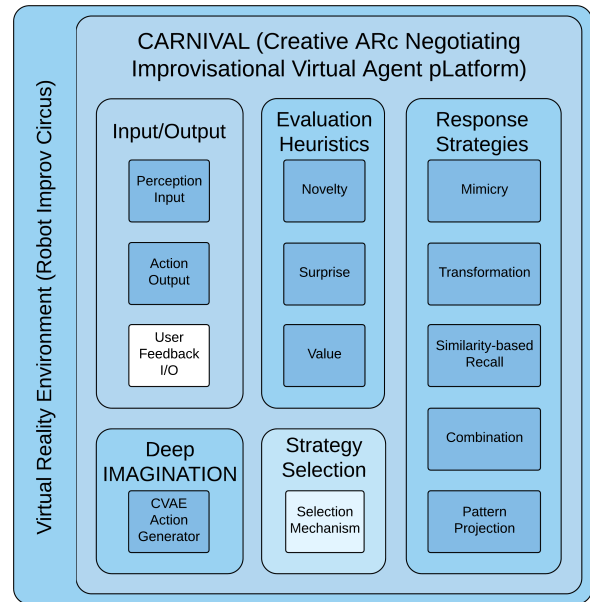
(b) The CARNIVAL agent architecture showing reasoning strategies, action generator, and evaluation heuristics. Partial shading shows future work.

Figure 3: Creative arc negotiation as implemented in the CARNIVAL architecture.

The Robot Improv Circus allows players to improvise with a virtual agent. The player, their virtual stage partner, and the audience are all humanoid robots limiting participant expectations for realism. During their turn, the player is given a prop and mimes a pretend action with it. After completing their action, they hit a buzzer to pass their prop to their agent partner. The agent improvises actions with the same prop before hitting the buzzer and passing it back to the player. The agent announces what it is trying to do using text-to-speech audio and a speech bubble. In Fig. 1, the agent is trying to mime looking through a kaleidoscope with a long thin prop. The game ends after a fixed number of turns. The improvisation is showcased to external spectators outside the installation using a virtual audience view.

4 CREATIVE ARC NEGOTIATION IN CARNIVAL

Creative arc negotiation for playing the Props game in the Robot Improv Circus is implemented in the CARNIVAL (Creative ARc Negotiating Improvisational Virtual Agent pLatform) agent architecture [13]. The term 'negotiation' is used because the agent considers and weights estimates of novelty, unexpectedness, and quality for both the human participant's actions and its own while following a given creative arc. To explain how this process works, we use a running example where the agent has been given a rising creative arc, with novelty, unexpectedness, and quality targets rising over time, and a long, thin, cylindrical prop. CARNIVAL then does the following process. First, given a creative arc to follow (the

rising arc), the agent uses a set of improvisational reasoning strategies adapted from previous literature [12] to connect its actions to what the human did before and guide its search over potential actions (actions for the long, thin prop). These strategies include *mimicry* (imitating a human action), *transformation* (transforming various elements of a human action before performance), *combination* (combining multiple human actions together), *similarity-based recall* (recalling the nearest or farthest action to a recent human action from episodic memory), and *pattern projection* (looking for temporal patterns between agent-human action pairs and applying that to future human actions).

All reasoning strategies are executed in parallel to generate multiple candidate actions. Actions are generated using a deep generative model called DeepIMAGINATION (Deep IMprovised Action Generation through INteractive Affordance-based exploratION) [10]. DeepIMAGINATION is a conditional variational autoencoder (CVAE) [35] that generates mimed gestures conditioned on the physical affordances [6] of the agent's current prop (different actions are generated for a long, thin, cylinder vs. a short, thick, pyramid). During training, a data set of high-dimensional human gestures captured in VR and attributes of props used to enact them are fed to the CVAE. The model learns to map gestures to points within a low-dimensional latent space (it learns to group similar gestures together for props with similar physical attributes). During improvisation, the agent can sample from the model's latent space to generate potential actions to consider, correctly conditioned on the physical affordances of the agent's current prop ('stabbing' is generated for a long, thin, cylinder instead of 'sipping tea' for a short,

thick, pyramid). The reasoning strategies mentioned previously, systematically sample from the model's latent space to generate candidate actions.

Each candidate generated action is then evaluated by heuristics that measure its *novelty*, *unexpectedness*, and *quality* to locate that action in the agent's creative space. So using a long, thin, cylinder as a sword might be low in novelty and surprise, but using it as your unicorn horn might be higher in both. Human actions are also evaluated similarly for the agent to reason about. The evaluation models act on the gestural (the representation of agent movement, say when swinging a 'sword') and semantic (natural language labels for pretend actions and objects, like the labels 'swing' and 'sword') contents of the perceived action. Novelty is measured by the agent as the aggregated distance between a perceived action and other comparable experiences that the agent experienced (the difference between swinging a sword and similar swinging actions). The agent measures unexpectedness (as a proxy for surprise) using a combination of Bayesian Surprise [8] and direct computation of deviation from expectation [19]. Unexpectedness is computed for both the choice of pretend object and pretend action (how unexpected is the choice of 'sword' given a long, thin, cylinder and the choice of 'swinging' given a pretend sword). These properties are termed *object surprise* and *action surprise* respectively. Quality (as a proxy for value) is then computed using smoothness of motion (how smoothly a sword was swung) and 'recognisability' of the gestural and semantic components of the action respectively (how uniquely representative was the chosen action given other potential choices). Finally, each evaluated action and the agent's estimate of the human participant's last action are aggregated to find its effective location in the creative space. The nearest of these actions to the next point on the target creative arc (so for a rising arc, the next action that is more novel, more unexpected, and higher quality) is chosen for performance and played back.

5 METHODOLOGY

We used the following research questions to evaluate the effect of creative arc negotiation for action selection on user experience in the Robot Improv Circus.

RQ1: Can observers and improvisers identify a creative arc when an agent used it for decision-making?

RQ2: Does creative arc negotiation result in better observer and improviser experiences compared to random selection?

RQ1 addresses whether the creative arc used by the agent is evident in the improvised performance regardless of whether subjects enjoyed the performance. Recognition of the different creative arcs would validate the agent's ability to create qualitatively different experiences mirroring the arcs. RQ2 investigates whether subjects prefer improvisation with creative arc negotiation to random action selection in terms of enjoyment, creativity, and logical coherence.

5.1 Observer Recognition of Creative Arcs (RQ1)

We started our evaluation of RQ1 with a survey-driven, observer-rating study. Observers were considered separately from improvisers due to their crucial but differing role in improvised performing arts like improv theatre. A hundred non-expert raters on Amazon

Mechanical Turk [34] were each asked to watch videos of three different improvised sessions between a researcher and the agent in the Robot Improv Circus installation (see Fig. 1b). For each video, they were then asked questions to choose whether a specific property of the performance was rising, falling, or level. In this configuration, random guessing would have a one in three (33.33%) chance of being correct. The alternative hypotheses (H_A) stated that for each evaluated property and a given creative arc, the proportion of subjects identifying the correct arc would differ significantly from those incorrectly identifying other arcs. The null hypotheses (H_0) stated that for each evaluated property, there would be no significant differences.

Three creative arcs were used in the recorded performances for comparison — *rising*, *falling*, and *level* arcs. Values for novelty, unexpectedness, and quality (as defined in section 4) along these arcs, increased uniformly, decreased uniformly, and stayed the same respectively. For a rising arc, this meant that the agent would try to increase the novelty, unexpectedness, and quality of its actions throughout the performance.

The agent used one of the three creative arcs described above in each video. Subjects were asked to determine whether a rising, falling, or level arc was being used by the agent in that video in terms of the *novelty*, *object surprise*, *action surprise*, *quality*, and *user-defined creativity* of the performance (defined previously in section 4). All subjects were given the definition of each property alongside the questions except for user-defined creativity. Subjects were asked to define creativity before the rating task started and were reminded to use that definition whenever user-defined creativity was evaluated.

5.2 Effect of Creative Arc Negotiation on Observer Experience (RQ2)

We used another survey-driven, observer-rating study to evaluate RQ2. The creative arc negotiation agent used either *rising*, *falling*, or *level* creative arcs, exactly as in the previous study (section 5.1). Our baseline sampled actions uniformly at random from the agent's latent space (section 4), ensuring meaningfully generated random actions that were still appropriate for the given prop. We chose this baseline due to a lack of other established action selection mechanisms implied by the improvisational action selection problem.

Subjects for the study consisted of 100 non-expert raters on Amazon Mechanical Turk [34]. They were asked to watch videos of two different sessions between a researcher and the agent using either creative arc negotiation or randomly sampled actions to improvise in the Robot Improv Circus. For each video, they were then asked to choose whether they preferred the one on the left or the one on the right in a forced-choice configuration based on different perceived properties of the performance. In this configuration, each video had a random probability of being selected 50% of the time. The different qualities they were asked to compare were *enjoyment*, *user-defined creativity of the agent*, and *logical coherence*. At the start of the study, participants were made to define creativity and reminded to use that definition during the task.

The initial experiment was also repeated with an identical methodology using videos with just the agent's turns spliced together from the original performance videos (the researcher's actions were

removed). This was done to mitigate any potential bias in the results emerging from the human's actions, i.e., in case their actions contributed positively or negatively to observer preferences. The sample size was also increased to 120 participants.

5.3 Creative Arc Negotiation and Improviser Experience (RQ1 + RQ2)

We conducted an improviser-rating, laboratory study with non-experts to get quantitative and qualitative feedback about the improviser experience of interacting with an agent using creative arc negotiation. The in-person experiment combined our methodologies from the two observer experiments and asked improvisers to identify creative arcs (RQ1) and compare experiences between creative arcs or random selection (RQ2).

Eighteen participants were recruited for the initial study in two batches (six and twelve subjects) from a non-expert student population. The number of responses obtained per question were either twelve or eighteen, since additional questions were asked of the second batch of participants. No other differences in methodology existed between these two populations, and the number of responses for each question is noted when reporting results.

Participants were first given an opportunity to get familiar with how to use the VR system and the specific installation through a tutorial and a set of trial rounds for the installation. Participants were next placed into one of 3 groups at random and continued on to complete two study tasks. Finally, the study concluded after participants were debriefed and compensated for their participation.

The first experimental task was a comparison between creative arc negotiation and random action selection (RQ2). Participants were assigned to 3 groups. Each group had 1 of 3 possible creative arcs and 1 no arc session. The ordering for conditions within each group was randomized across participants. Each participant was asked to perform two sessions of improvisation with the agent. In each session, the agent used a different action selection mechanism according to the participant's assigned group. After improvising with the agent twice, the participant was asked to compare the two sessions through a survey followed by a semi-structured interview.

The session comparison questionnaire for these tasks asked the following two to three questions (depending on the study batch).

(1) Which of the sessions did you enjoy more?
(2) In which of the sessions would you say your partner was more creative overall?
(3) Which of the two sessions would you say was more logical overall?

These questions received 18, 18, and 12 responses. For the second question, participants were asked to reflect on their own definition of creativity before completing this questionnaire. For all questions, participants could select between the options — session one, session two, both equally, and neither. During the semi-structured interview, participants were asked questions to clarify their definition of creativity used in the questionnaire, memorable reasons or examples of interactions that led to their responses, and other reasons why they preferred one session over the other. Participants were also asked for open-ended feedback on the interaction, experience, or any other aspect of the sessions.

The second experimental task was creative arc recognition (RQ1). Participants were assigned to 3 groups, 1 for each pairing of rising, falling, and level creative arcs. The ordering of conditions was randomised within each group. Participants performed two sessions of improvisation with the agent, answering questions after each session.

Each session was evaluated with a questionnaire. It asked participants whether the novelty, object surprise, action surprise, and user-defined creativity increased, decreased, or stayed the same over time. Definitions for each property (except user-defined creativity) were presented to them alongside each question as defined in section 4.

6 RESULTS

We present our results on the effect of creative arc negotiation on observer and improviser experiences in the Robot Improv Circus in this section. We start with the results for each study (observer creative arc recognition, observer creative arc comparison, and participant creative arc recognition with comparison) and then discuss their implications in the next section.

6.1 Observer Recognition of Creative Arcs (RQ1)

The relative percentages of participants who correctly (and incorrectly) identified each property of the performance as rising, falling, or level are presented in Table 1. Note that random guessing would score approximately 33.33%. The total parameter reports accuracy over all other parameters combined.

A chi-square goodness of fit test was performed to evaluate whether there were significant results between correctly vs. incorrectly identifying the direction of the arc for the given property. The null hypotheses for the five questions were that there was no significant difference between the distributions of responses identifying the arc for each property. The alternate hypotheses stated that significant differences did exist between these distributions. The results for each rated property of the session from the chi-square goodness of fit test are presented in table 2.

The creative arc recognition accuracies and the statistical hypotheses testing showed that for *novelty*, *object surprise*, and *creativity*, performances with *rising* and *falling* arcs could be identified reliably. Surprisingly, it also showed that recognition accuracies for all significantly differing properties of *level* arcs were consistently as bad as random guessing.

The chi-square goodness of fit results showed medium and large effects for a subset of significant results. The results indicate that observers reliably noticed the intended changes in those properties during the performance. Most importantly, these results conclude that **for observers, most queried properties defining our creative space were reliably identifiable with rising arcs, less reliably identifiable with falling arcs, and difficult to identify for level arcs.**

6.2 Effect of Creative Arc Negotiation on Observer Experience (RQ2)

The percentage of participants who preferred creative arc negotiation over no arc random sampling for each rated property of

Table 1: Relative recognition percentages between arc types in creative arc identification task (RQ1). Bold is higher between pairs.

	Rising		Falling		Level	
	Correct	**Incorrect**	**Correct**	**Incorrect**	**Correct**	**Incorrect**
Novelty	57.14%	42.86%	37.14%	**62.86%**	20.95%	**79.05%**
Object Surprise	53.33%	46.67%	44.76%	**55.24%**	34.29%	**65.71%**
Action Surprise	47.12%	**52.88%**	37.14%	**62.86%**	42.86%	**57.14%**
Quality	73.08%	26.92%	**61.90%**	38.10%	**60.00%**	40.00%
Creativity	51.43%	48.57%	43.81%	**56.19%**	33.33%	**66.67%**

Table 2: Chi-square goodness of fit for creative arc identification task arcs (RQ1). Bold significant at $p < 0.05$. ϕ_c is effect size.

	Rising			Falling			Level		
	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c
Novelty	26.79	**$< 10^{-5}$**	**0.505**	0.69	0.40763	0.081	7.24	**0.0071**	**0.263**
Object Surprise	18.90	**$< 10^{-5}$**	**0.424**	6.17	**0.013**	**0.242**	0.04	0.836	0.020
Action Surprise	8.89	**0.0029**	**0.292**	0.69	0.4076	0.081	4.29	**0.0384**	**0.202**
Quality	73.92	**$< 10^{-5}$**	**0.843**	38.57	**$< 10^{-5}$**	**0.606**	33.60	**$< 10^{-5}$**	**0.566**
Creativity	15.47	**0.00008**	**0.384**	5.19	**0.0228**	**0.222**	0	1	0

the performance are presented in table 3. The properties that participants were asked about included which session they enjoyed more (*enjoyment*), in which session the agent seemed more creative (*Agent Creativity*), and which session seemed more logically coherent (*Coherence*). Note that random guessing would score 50% in this task.

A chi-square goodness of fit test was performed on all the queried properties to evaluate whether there were significant differences in observer preferences between creative arc and no arc conditions for each arc. The null hypotheses for comparing each queried property were that there were no significant differences between the distributions of responses for each arc to the distribution of expected outcomes in each case. The alternate hypotheses stated that significant differences did exist between these distributions. The results for each rated property of the session from the chi-square goodness of fit test are presented in table 4.

The results of this experiment suggest that there were **significant, reliably detectable preferences for *rising* and *falling* creative arc negotiation performances in comparison with a random sampling baseline**, at least for observers viewing videos of performances. *Level* arcs significantly and reliably were preferred less than the random selection baseline. All properties (*enjoyment*, *agent creativity*, and *coherence*) showed significant results with effect sizes ranging from small to large. The effects for rising and falling arcs (with the effect stronger in general for rising arcs) showed that coherence was the most improved, with agent creativity and enjoyment following closely behind.

The results from the repeated study with just the agent's actions spliced together in the observer's video were analyzed exactly the same way as the previous analysis. The results for recognition accuracies across arcs can be seen in table 5. After performing statistical significance testing, the results can be seen in table 6.

The results for the repeated observer study with footage of just the agent taking its turns **showed an even stronger effect in the same direction as the version with the human performer.** This allowed us to remove the effect of the human on the observed results. It also allowed us to address any potential concerns about researcher bias in terms of implicitly shaping the videos for evaluation. This is a valid concern since it is a co-creative performance with creative responsibilities falling on the shoulders of both human and computer improviser. It would be natural for there to be researcher bias or error while constructing the comparison videos. However, the results from the repeated iteration of the study lay any such concerns to rest and improve on the previous results in terms of effect size and increased preference for the creative arc negotiation versions of the system.

6.3 Creative Arc Negotiation and Improviser Experience (RQ1 + RQ2)

The results for the creative arc comparison and identification tasks are presented in this section.

6.3.1 Creative Arc Comparison (RQ2). The results from the questionnaire for the creative arc comparison study task are summarized and presented in Tables 7 and 8. Table 7 summarizes the relative differences in four possible preferences (creative arc, no arc, both, or neither) between the two conditions compared in the task (creative arc and no arc). Table 8 shows the result of performing a chi-square goodness of fit test on the combined data for comparing creative arc sessions against no arc sessions. The null hypotheses for each queried property compared was that there were no significant differences between the distribution of preferences for creative arc negotiation sessions to the distribution of expected outcomes. The alternate hypotheses for these properties were that there were significant differences for these distributions. For both analyses of

Table 3: Relative preferences between 'arc' and 'no arc' conditions in creative arc comparison task (RQ2). Bold is higher in pair.

	Rising		Falling		Level	
	Arc	**No Arc**	**Arc**	**No Arc**	**Arc**	**No Arc**
Enjoyment	**63.46%**	36.54%	**62.50%**	37.50%	28.57%	**71.43%**
Agent Creativity	**64.08%**	35.92%	**56.73%**	43.27%	31.73%	**68.27%**
Coherence	**80.58%**	19.42%	**75.96%**	24.04%	24.04%	**75.96%**

Table 4: Chi-square goodness of fit for creative arc comparison task arcs (RQ2). Bold significant at $p < 0.05$. ϕ_c is effect size.

	Rising			Falling			Level		
	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c
Enjoyment	7.54	**0.00604**	0.269	6.5	**0.01079**	0.250	19.29	$< 10^{-5}$	0.429
Agent Creativity	8.17	**0.00427**	0.282	1.89	0.013	0.135	13.89	**0.00019**	0.365
Coherence	38.53	$< 10^{-5}$	0.612	28.04	$< 10^{-5}$	0.519	28.04	$< 10^{-5}$	0.519

Table 5: Relative preferences between an arc condition and a no arc condition in creative arc comparison task (RQ2) with only agent turns (no human turns). Bold is higher between pairs. Surprisingly, preferences are stronger without human turns.

	Rising		Falling		Level	
	Arc	**No Arc**	**Arc**	**No Arc**	**Arc**	**No Arc**
Enjoyment	**86.67%**	13.33%	**70.83%**	29.17%	23.53%	**76.47%**
Agent Creativity	**73.33%**	26.67%	**63.03%**	36.97%	25.21%	**74.79%**
Coherence	**93.33%**	6.67%	**73.33%**	26.67%	16.81%	**83.19%**

Table 6: Chi-square goodness of fit for creative arc comparison task (RQ2) with only agent turns (no human turns). Bold significant at $p < 0.05$. ϕ_c is effect size. Results are stronger without human turns.

	Rising			Falling			Level		
	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c	χ^2	**p**	ϕ_c
Enjoyment	64.53	$< 10^{-5}$	0.733	20.83	$< 10^{-5}$	0.417	33.35	$< 10^{-5}$	0.529
Agent Creativity	26.13	$< 10^{-5}$	0.467	8.08	**0.00449**	0.261	29.25	$< 10^{-5}$	0.496
Coherence	90.13	$< 10^{-5}$	0.867	26.13	$< 10^{-5}$	0.467	52.45	$< 10^{-5}$	0.664

the creative arc comparison task, our sample size was too small to split the different comparisons by arc types. Therefore, we did not separately compare preferences for each arc type against no arc sessions.

The semi-structured interview data contained explanations for subject questionnaire choices, definitions for creativity, other memorable aspects of the sessions, and feedback about the experience. This data was transcribed and analysed to surface key themes and trends among the subjects' responses. The results offer additional support for the effectiveness of creative arc negotiation on improviser experience.

Participant explanations for why they enjoyed one session over the other included several distinct themes. Two (of 5) subjects who said they enjoyed both sessions equally did so because of how engaging and immersive they found the physicality of interaction in VR, making it harder to notice any differences in the agent's

behaviour. The majority of subjects who said they enjoyed the no arc session more (4 of 6), also said that they enjoyed it more because the prop they received in that session was easier to map to different objects for pretending with. This is encouraging, since their preference was not based on the agent's performance. A third of all subjects (6 of 18) described the novelty of agent actions as their main reason for their response. However, a greater number (and proportion) of subjects who enjoyed the creative arc session more provided this reason compared to no arc-preferring subjects (4 of 7 vs. 2 of 6). This also suggests that greater novelty can be attained when novelty is part of the search criteria than with random selection.

Participants also provided the definitions of creativity they used and reasons for why they judged agents in one session more creative than the other. The vast majority of participants (16 of 17 who gave their definition) considered novelty + surprise as main factors in

Table 7: Relative preferences for 'arc', 'no arc', 'both', or 'neither' between a creative arc session and a 'no arc' (random action selection) session in the participant-rating creative arc comparison task (RQ2). Bold is highest for the row. N is sample size.

	Creative Arc	No Arc	Both	Neither	N
Enjoyment	**38.89%**	33.33%	27.78%	0%	18
Agent Creativity	**55.56%**	27.78%	5.56%	11.11%	18
Coherence	**58.33%**	8.33%	25%	8.33%	12

Table 8: Chi-square goodness of fit between 'arc' and 'no arc' sessions for the participant-rating creative arc comparison task (RQ2). Bold significant at $p < 0.05$. ϕ_c and $\tilde{\phi}_c$ are effect size and adjusted effect size respectively. N is sample size.

	χ^2	p	ϕ_c	$\tilde{\phi}_c$	N
Enjoyment	6.44	0.09188	0.598	0.439	18
Agent Creativity	**10.89**	**0.01234**	**0.778**	**0.675**	18
Coherence	**8**	**0.04601**	**0.816**	**0.658**	12

their definition. Some subjects seemed to combine concepts like novelty and surprise in their responses, so the two concepts are combined together here. Similarly, the vast majority of subjects chose one session over the other in terms of agent creativity because of a perceived difference in diversity or novelty (8 of 10 for creative arc sessions and 4 of 5 for no arc sessions). Four subjects overall considered it more creative when ideas were generated for "harder" props and 2 of 10 subjects chose the creative arc session as a result.

Subjects attributed the logical coherence of sessions (or the lack thereof) to different reasons. Three (of 7) subjects who chose the creative arc session highlighted a perceived narrative or story-like sequence to the agent's actions exemplified by the quote, "There was more of a sequence of events. . . . like a beginning, middle, and end to the story . . . it felt more story-like." Two subjects thought neither agent was coherent with 1 saying, "Both were equally illogical." One (of 7) who chose the creative arc session mentioned how the agent paid attention to their actions.

The statistically significant results from the chi-square analysis for the session comparison responses, large effect sizes, the direction of preferences, and subjects' responses in interviews showed promise for the effect of creative arc negotiation on improviser experience, at least compared to a random action selection baseline. **Subjects reliably and strongly preferred creative arc negotiation when evaluating agent creativity and logical coherence between the two types of sessions.** User enjoyment did not differ significantly between the two conditions. However, further study is required to rule out an effect on user enjoyment, since 6 of 11 subjects who enjoyed arc negotiation less, described coincidental or task-specific differences unrelated to the agent's behaviour as reasons for their choice.

6.3.2 Creative Arc Identification (RQ1). The recognition percentages were combined across all creative arc types for each queried property in the session creative arc identification questionnaire. These results showed performances similar to random guessing (at around 33.33%). These results were also not significant in a chi-square goodness of fit test. This is a negative result since improvisers could not differentiate between creative arcs driving agent

actions. However, it is important to note that our experimental design used creative arc identification tasks to understand whether subjects could identify trends in their experience that the agent was attempting to modulate regardless of whether they preferred experiences with that form of action selection. We discuss potential reasons for these results in section 7. These initial results for creative arc identification will also be expanded and reviewed in the future after increasing the sample size to get higher confidence results.

7 DISCUSSION

The evaluation studies in this work aimed to grasp the effect of creative arc negotiation on observer and improviser experience. Could they identify what kind of creative arc an agent used in a performance (RQ1)? Would they prefer improvisation sessions guided by creative arc negotiation over an alternative (RQ2)?

The results from the two observer studies provided strong positive evidence that observers both identified and preferred creative arcs, though this phenomenon was applicable to different effect sizes based on the arc. The result that observers preferred performances guided by creative arc negotiation in terms of enjoyment, agent creativity, and logical coherence, empirically demonstrates that **creative arc negotiation addresses the improvisational action selection problem for observers of improvised performances.** The existence of these preferences makes sense given the body of literature on the presence of specific sets of arcs across narratives for dramatic tension and plot [1, 5, 36] or character affect [29]. The result that observers had a significant preference for random sampling over level arcs was unexpected and requires more study. Perhaps random selection resulted in more novelty. The laboratory study did indicate that novelty was key for perceptions of agent creativity and subject preferences.

The results from the in-person laboratory study indicated that **improvisers also preferred creative arc negotiation for agent creativity and coherence though not necessarily for enjoyment.** The interviews did partially explain why enjoyment was

evenly distributed but further study is needed. The negative results for creative arc identification warrant deeper review. Evidence [15, 16] suggests that interactors in the midst of an ephemeral interactive experience have trouble keeping track of longer-term effects. Kelso et al. [15] interpret this positively, since interactive narratives then do not need strong narrative coherence, unlike other forms of narrative, if participants cannot keep track of these longer-term links. However, their finding also implies that demonstrating strong trends in the user's experience based on longer-term effects would be harder in these experiences.

Our empirical evaluation above provides initial evidence that creative arc negotiation can be successfully used for improvisational action selection. Further evaluation could consider longer, non-monotonic creative arcs and their use in performances in the wild. Future work could also investigate creative arcs as a technique for negotiating between authorial intent and agent autonomy – a general challenge in many co-creative and human-AI domains [11, 30].

8 CONCLUSION

This article describes creative arc negotiation in the CARNIVAL architecture as a solution to the improvisational action selection problem. We also briefly describe the Props game in the Robot Improv Circus installation as a domain for studying creative arc negotiation. Finally, we contribute three experiments to understand the effect of creative arc negotiation in the Robot Improv Circus on the experience of observers and participating improvisers.

The results from the creative arc identification and comparison studies for observers showed that, depending on the specific arc, they could identify trends in the performances that corresponded to the creative arc used by the agent. The experiments also showed their preference for action selection using creative arc negotiation over a random sampling baseline in terms of perceived enjoyment, agent creativity, and logical coherence. Observer preferences were even stronger when the human partner's actions were removed from the tasks, and only the agent's actions were evaluated.

Creative arc identification was not successful in our small-scale laboratory study for participating improvisers, unfortunately, and we discussed why this might be the case. More positively, evidence from the creative arc comparison task showed that, depending on the specific arc, perceptions of agent creativity and logical coherence were significantly higher (enjoyment required more study for conclusive results) for improvised performances with creative arc negotiation compared to our baseline. Therefore, **creative arc negotiation can be successfully used by improvisational agents for action selection and performances created this way are empirically preferred over our baseline by both observers and participating improvisers.**

REFERENCES

[1] Aristotle. and S. H Butcher. 1969. *Poetics*. Hill and Wang.
[2] Joseph Bates. 1992. Virtual reality, art, and entertainment. *Presence: Teleoperators & Virtual Environments* 1, 1 (1992), 133–138.
[3] M. A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.
[4] António Brisson, Brian Magerko, and Ana Paiva. 2011. A computational model for finding the tilt in an improvised scene. In *International Conference on Interactive Digital Storytelling*. Springer, 158–163.
[5] Gustav Freytag. 1896. *Freytag's technique of the drama: an exposition of dramatic composition and art*. Scholarly Press.
[6] James J Gibson. 1979. *The ecological approach to visual perception*. Psychology Press.
[7] Daniele Gravina, Antonios Liapis, and Georgios Yannakakis. 2016. Surprise search: Beyond objectives and novelty. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*. ACM, 677–684.
[8] Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research* 49, 10 (2009), 1295–1306.
[9] Mikhail Jacob. 2019. *Improvisational artificial intelligence for embodied co-creativity*. Ph.D. Dissertation. Georgia Institute of Technology.
[10] Mikhail Jacob, Prabhav Chawla, Lauren Douglas, Ziming He, Jason Lee, Tanuja Sawant, and Brian Magerko. 2019. Affordance-based generation of pretend object interaction variants for human-computer improvisational theater. In *Proceedings of the 10th International Conference on Computational Creativity (ICCC)*.
[11] Mikhail Jacob, Sam Devlin, and Katja Hofmann. 2020. "It's Unwieldy and It Takes a Lot of Time"—Challenges and Opportunities for Creating Agents in Commercial Games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. 88–94.
[12] Mikhail Jacob and Brian Magerko. 2015. Interaction-based Authoring for Scalable Co-creative Agents. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*. Provo, UT.
[13] Mikhail Jacob and Brian Magerko. 2018. Creative Arcs In Improvised Human-Computer Embodied Performances. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*. ACM, 62.
[14] P.N. Johnson-Laird. 2002. How Jazz Musicians Improvise. *Music Perception* 19, 3 (2002), 415–442.
[15] M. T Kelso, P. Weyhrauch, and J. Bates. 1993. Dramatic presence. *Presence: The Journal of Teleoperators and Virtual Environments* 2, 1 (1993), 1–15.
[16] Brenda Laurel. 1991. Computers as Theatre Reading. *Mas: Addison-Wesley Publishing Company* (1991).
[17] Joel Lehman and Kenneth O Stanley. 2011. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation* 19, 2 (2011), 189–223.
[18] Duri Long and Brian Magerko. 2020. Why don't computers improvise with us?. In *CHI 2020 Workshop on Artificial Intelligence for HCI: A Modern Approach (AI4HCI)*.
[19] Mary Lou Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*. Desire Network, 22–28.
[20] Lara J Martin, Brent Harrison, and Mark O Riedl. 2016. Improvisational Computational Storytelling in Open Worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*. Springer, 73–84.
[21] Michael Mateas and Andrew Stern. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference*, Vol. 2. 4–8.
[22] Kory W Mathewson and Piotr Mirowski. 2017. Improvised theatre alongside artificial intelligences. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.
[23] David Mendonça and William A. Wallace. 2004. Cognition in Jazz Improvisation: An Exploratory Study. Chicago, IL.
[24] Kathryn E Merrick and Mary Lou Maher. 2009. *Motivated reinforcement learning: curious characters for multiuser games*. Springer Science & Business Media.
[25] Brian O'Neill, Andreya Piplica, Daniel Fuller, and Brian Magerko. 2011. A knowledge-based framework for the collaborative improvisation of scene introductions. In *Proceedings of the 4th International Conference on Interactive Digital Storytelling*, Vol. 7069 LNCS. Vancouver, Canada, 85–96.
[26] Andreya Piplica, Christopher Deleon, and Brian Magerko. 2012. Full-body gesture interaction with improvisational narrative agents. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7502 LNAI. 514–516.
[27] Julie Porteous, Jonathan Teutenberg, David Pizzi, and Marc Cavazza. 2011. Visual programming of plan dynamics using constraints and landmarks. In *Twenty-First International Conference on Automated Planning and Scheduling*.
[28] Jeff Pressing. 1984. Cognitive processes in improvisation. In *Advances in Psychology*. Vol. 19. Elsevier, 345–363.
[29] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (2016), 31.

[30] Mark Owen Riedl and Vadim Bulitko. 2013. Interactive narrative: An intelligent systems approach. *Ai Magazine* 34, 1 (2013), 67–67.

[31] David L Roberts and Charles L Isbell. 2008. A survey and qualitative analysis of recent advances in drama management. *International Transactions on Systems Science and Applications, Special Issue on Agent Based Systems for Human Learning* 4, 2 (2008), 61–75.

[32] Christoph Salge and Christian Guckelsberger. 2016. Does empowerment maximisation allow for enactive artificial agents?. In *Proceedings of the Artificial Life Conference 2016 13*. MIT Press, 704–711.

[33] Jürgen Schmidhuber. 2006. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science* 18, 2 (2006), 173–187.

[34] Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs* 85, 1 (2018), 140–156.

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*. 3483–3491.

[36] Kurt Vonnegut. 2005. At the Blackboard. *Lapham's Quaterly* (2005).