Five Years after the First Theatre Play Generated by GPT-2

Rudolf Rosa*

Charles University, Faculty of Mathematics and Physics, Malostranské náměstí 25, 118 00 Praha, Czech Republic

Abstract

In 2021, we staged the first theatre play generated by a large language model with minimal human intervention. It was the first such project in the world, dealing with numerous then-open questions of integration of AI technology into theatrical practice.

Five years later, we look back on the project, putting it into context of the situation back then as well as the development since then. We discuss various aspects of the process of generating the script, staging the play, and the reception of the play by the theatre professionals as well as the audiences.

Keywords

Natural Language Processing, theatre, artificial intelligence, literary NLP

1. Introduction

In 2021, we staged AI: When a Robot Writes a Play, the world's first full-length theatre play with a mostly AI-generated script. In this paper, we reflect back on the project with a distance of five years, allowing us to view the project in the context of further development in the field. More broadly, the project serves as a case study around which we center our observations on employment of AI in art, with a special focus on theatre.

1.1. The THEaiTRE Project

THEaiTRE[1] [1, 2] was our interdisciplinary project directly combining theatre and science, with the main aim of producing and staging the first computer-generated theatre play. The project started in 2019, producing its first theatre play in 2021, and concluding with a second play in 2022.

The initial motivation behind the project was to celebrate the 100^{th} anniversary of Karel Čapek's theatre play *R.U.R.* [3], for which the word *robot* was invented and which was a play about robots. Our idea was to turn R.U.R. around by having a "robot" (actually, artificial intelligence) write a play about humans.

Within the project, we designed an interactive theatre script generation tool *THEaiTRobot* [4], and used it to produce scripts for two theatre plays, *AI: When a Robot Writes a Play* [5, 6] and *Permeation* [7]. For both of the plays, 90% of the script was generated automatically using THEaiTRobot; the remaining 10% are manually written parts provided as inputs (starting prompts) for the generation process and other manual post-edits of the scripts. The plays were then professionally staged, rehearsed and premiered in theatres. The only involvement of AI was in generating the dialogues of the characters; everything else, from stage directions to costumes and music, was done manually.

The goal of the project was to explore the potential of current artificial intelligence techniques to be incorporated into theatre practice, and to directly confront the general public with the outcome while explaining the process behind the creation of the play and thus educating the audience about the current

7th International Conference on Creative\Media/Technologies (IConCMT), November 26-27, 2025, University of Applied Sciences St. Pölten, Austria

*Corresponding author.

rosa@ufal.mff.cuni.cz (R. Rosa)

https://ufal.mff.cuni.cz/rudolf-rosa (R. Rosa)

0000-0003-4908-6127 (R. Rosa)

[1] https://www.theaitre.com/

state and capabilities of the techniques used. This immersive experience sparked a lot of interest among theatre experts, artificial intelligence enthusiasts, as well as the general public [8, 9, 10, 11, 12, 13, 14].

Our goal was to demystify AI, to explain the rather simple basic principles behind the machine learning based on textual training data, and to bring the AI closer to people. We tried to enable the public not to either glorify AI or fear AI irrationally, but rather understand its basic principles and thus set reasonable expectations and precautions, making it clear that AI should be taken with caution but can be very useful for some tasks in practice.

In the process, we faced numerous then-open questions of LLM training and deployment, of ways of interacting with LLMs, of the potential of incorporating AI technology into art, of authorship and copyright, as well as of reception by the audiences; most of these questions were rather niche at that time. Since the project, the AI technologies as well as these questions have become much more prominent. We reflect on the process in this paper.

1.2. Goals and Methodology of this Paper

The main goal of this paper is to use the THEaiTRE project as a case study for a more general reflection on the development and current state of the views and employment of AI in theatrical and other art. The 5-year distance allows us to view the field from a broader perspective, observing the emergence and development of various problems, concepts, views and methods. Through linking our experiences to further development in the field, as well as comparing to other existing projects in the field, we contribute to current discussions on AI in art. Thanks to the temporal distanciation, we can view the current development through the lens of the past, and to view the past through the lens of the present situation, thus allowing us to gain a more abstract and comprehensive perspective.

We start by reviewing related projects that engage AI in theatrical and scripted settings. We position them against our project and put them into broader context.

Next, we describe and discuss the technological aspects of the THEaiTRE project, putting them into context of contemporary and subsequent state and development of the technologies and approaches.

We then shift our focus to the artistic aspects of the project, paying considerable attention to the reception of the project by its members, the artistic community, and the audiences, and put the views of the project into the context of current views of employment of AI in artistic practice.

We conclude by summarizing our main findings and observations.

2. Related Projects

We review several notable projects involving theatre and/or scripts and artificial intelligence that preceded or were contemporary with our project. Such projects started to appear more frequently from the year 2016 on, which is around the time when LSTM-based neural language models became commonly used, considerably improving over previous approaches [15].

In our project, we opted for maximum transparency, making the THEaiTRobot tool publicly available online, releasing its source codes, describing the design of the system in several scientific papers, and publishing the generated theatre scripts with all human interventions marked. In the related projects, this is often not the case, with many of the devised tools not being publicly available and with little transparency about the particularities of the exact design and usage of the tools; the situation is typically better with projects set within academia. Moreover, most of the projects use substantial human curation and/or only produce short generated scripts or human-AI interactions; our project stands out by having produced a full 1-hour play script with 90% generated content.

We can broadly divide the related projects into two groups – scripted performances, where the script is fully created beforehand; and improvised performances, which involve creating and developing the script live within the performance.

2.1. Scripted Performances

Our project falls into the category of scripted performances: the script of the theatre play was generated, edited, rehearsed, and performed, with only minimal changes to the script after the editing phase. Such an approach allows to use the text generation tools more directly, without the need of developing innovative interactive solutions to actively take part in the live show.

The the best of our knowledge, the first generative art project that utilized modern neural language models to generate a script for a performance was the musical **Beyond the Fence**, which premiered in 2016 and was labelled as the "first computer musical". The project combined several tools which had been developed by various authors and institutions. The *What-If Machine* ideation engine and the *PropperWryter* plot generator suggested a concept and a plot, the *Claude-Machine Schönbot* and the *FlowComposer* music generating programs generated musical scores, and *Clarissa the Cloud Lyricist* generated lyrics. The members of the project gathered suggestions from the automated tools, post-edited them heavily or took inspiration from them, and compiled all of the contents into a full-length musical. The authors have estimated that approximately 25% of the final script is computer-generated. The musical was staged and repeatedly performed in February and March 2016 at the Arts Theatre in London's West End. This project also stands out by being accompanied by a detailed published scientific paper describing the tools and the generation process, as well as evaluating the project results [16]. In our project, we aimed for a much higher amount of AI share in the final script, prioritizing the revelation of the current state of the technology over the quality of the resulting play itself. On the other hand, we only generated the dialogues, with all the staging, music and costumes done manually by theatrical experts.

As for the high share of generated content in the script, the first project with the same aims to ours was presumably the **Sunspring** (2016). To the best of our knowledge, its script was generated completely automatically, including stage directions, with no human intervention into the contents of the script. The script was then interpreted by the director and the actors, resulting in a short 9-minute experimental sci-fi movie (the actual part based on the generated script is 7 minutes long). The movie as well as the script are available online.[2] However, due to the used LSTM-based language model, the script is rather boring and non-sensical, which is why the movie itself is much shorter than our play; the technology was not yet in the state to produce a full-length script. Only with the replacement of recurrent networks by self-attentive networks through the Transformer architecture [17], it became possible to maintain consistency and sensibility at a larger scale, allowing us to generate a script for a one-hour play (even though individual scenes still needed to be produced independently). On the other hand, unlike our project, the Sunspring script contains not only generated dialogues but also stage directions (even less sensible than the dialogues). Based on that experience as well as our preliminary unsatisfactory experiments, we opted for generating only the dialogues, without stage directions and other metatexts.

While all of these projects constitute scripted performances, none of them constitute a theatre play per se, as is the case of our project. The first actual theatre play with a generated script was **Lifestyle of the Richard and Family**, which premiered in 2018. According to available information, the human (co-)author, Roslyn Helper, wrote the theatre script with the help of a next word suggestion tool.[3] The human author started each of the sentences in the script, but some of the sentences were finished using SwiftKey Note. This was a tool that, based on already written text, offered the user with 3 options for what the next word could be, allowing to keep iteratively accepting these suggestions until the sentence is finished. The project was devised and the resulting play was directed by Harriet Gillies. The play was performed at the Meat Market for Next Wave Festival in Melbourne in May 2018. A short video containing excerpts from the play as well as comments by the project team is available on Vimeo.[4] This is thus quite similar to our project in the format, process, and length, although again, the ratio of the generated content in the final script is much lower. Unfortunately, the available information about the

[2] https://www.thereforefilms.com/sunspring.html
[3] https://www.harrietgillies.com/lotraf
[4] https://vimeo.com/263802155

project is not very rich and does not allow for a more detailed comparison.

The THEaiTRE project was thus not novel in any single aspect alone, but its novelty lies in the combination of several of the demanding aspects (producing a full-length theatre play script with a large share of generated content), which only became possible thanks to the release of the GPT-2 language model [18], much more powerful than the previously available tools.

2.2. Improvised Theatre

A different branch to scripted theatre, and thus to our project, is improvised theatre, where the script exists only partially before the performance, and the performance is co-created and shaped live by the actors on stage, thus delivering a different show with each repeat. While we did not choose this path with our project, it is arguably a setting more suited for AI employment, as is demonstrated by the larger amount and variety of projects in this branch as well as their larger success. The improvised theatre projects show that AI can become an interesting novel component of the live show, thanks to its ability to produce text and other outputs quickly and interactively, thus engaging into the show as an active non-human participant. Improvised theatre relies on novel inputs being actively produced within the performance, typically by the actors and/or the audiences; existing projects show that AI, if used sensibly, can take on exactly this role.

Since 2016 until the present time, the leading force in improvised theatre with AI has been the **Improbotics** group[5] (formerly also as HumanMachine). It has brought together theatre performers and artificial intelligence researchers from various countries and continents and features a range of theatrical projects, both past as well as ongoing, which make use of automatically generated textual content (as well as other technology) combined with improvisation.

Some projects of the group employ very similar technology to the THEaiTRE project, utilizing the GPT family of generative neural language models. A prime example is their engagement of a "cyborg actor" who says and acts out lines that are generated in real-time by an on-stage operator running a GPT based system (originally, GPT-2 was used; but the setup updates as new GPT versions become available). The generated lines are sent through an automated text-to-speech system to an earpiece that the actor is wearing and spoken out loud by them. A second, fully human actor then improvises around what the cyborg says and does. The audience can then be subjected to a variation of the Turing test, trying to discern which of the actors is the cyborg.

The members of the group also frequently publish scientific papers describing their methods as well as discussing usage of AI in improvised theatre in general [19, 20, 21, 22]. The team has also already used machine translation live on stage [23], as well as incorporating live text-to-speech synthesis [24], both with automatically generated content as the input.

The **OK CARBON**[6] show, premiered in 2021, made the presence of the AI even more overt, by actively putting a computer system on stage as a visible active participant. The partially improvised revue show combines several parts, reflecting on and interacting with technology in various ways (including e.g. operating an automatic piano or burning of transistors). The title of the show and one of the parts of the performance is the inverse of the "OK, Google" prompt that activates the Google assistant to fulfill your requests. In the performance, it is the carbon-based human actors who are the servants, with an automated system giving them commands that the actors fulfill or questions that the actors answer.

This is yet another approach to human-machine interaction, where the machine is not an assistant or a partner, but it is rather the master of the humans. While this view is common in sci-fi literature and film, it only very rarely happens in practice; even though such cases exist even in the real world, such as autonomous taxi cars ordering human services to close their doors.[7]

In our project, it was always the human operator of the system who was in charge; although we aimed to minimize their agency as much as possible, thus leading to the AI system having more influence on

[5] https://improbotics.org/

[6] https://www.alfredvedvore.cz/cs/program/ok-carbon-170/

[7] https://www.cnbc.com/2026/02/12/waymo-is-paying-doordash-gig-workers-to-close-its-robotaxi-doors.html

the final script, but the human having both the initial and the final word.

2.3. Artistic Research

The THEaiTRE project, as well as all of the aforementioned projects, are primarily applied and practical, with the main aim of delivering a live performance. Even though they all do have some research aspects to them, their main aim is not research.

However, within the improvised AI theatre branch, the further development of the technology as well as the experiences from the practical projects have eventually led to establishment of several artistic research projects, which utilize AI in the theatrical setting with the primary aim of exploring explicit research questions.

In **The Answering Machine** (since 2022), Gunter Lösel combines his experience as an improvisation actor as well as a theatrical researcher into a unique artistic research project, using artistic skills and approaches to understand artificial intelligence models embedded into "social bots" [25]. Artificial intelligence models are hard to understand (often compared to "black boxes"), with a considerable amount of research being done in technical fields on interpreting and probing the models through technical means [26, 27], However, systems built on top of LLMs can exhibit social and human-like behaviour, suggesting that another possible way of trying to understand them may be through methods designed to be applied to humans. Through the method of applied anthropomorphization, the research team interacts with "social bots" on stage within improvised theatrical performances in various ways. The approach builds upon skills typical for theatrical professionals and psychologists of trying to understand someone (a person or theatrical character) by feeling into them. By trying to feel into the artificial persona, we may learn something about the underlying AI models, in a very human way which is very different from the typical technological interpretability approaches.

In another project, **The Feeling Machine** (since 2018), Manuel Flurin Hendry has been experimenting with artificial intelligence tools to work with emotions [28].[8] While LLMs are more and more capable in working with text, Hendry notes that there is much more to human communication than the words. In theatre, working with feelings and emotions is of equal importance, if not more important than working with words and meanings. At the same time, emotions are often mentioned as one of the typically human traits that is not yet fully accessible to machines. Therefore, Hendry and his team focus on creating a virtual avatar in the physical space that should be able to express emotions by applying traditional acting methods. Similarly to our project, the main goal is to allow for the general public to encounter and interact with this virtual actor, so that the audiences can see and feel the current state of technology, allowing them to gain practical insights and understanding. At the same time, the experience is reported to be entertaining yet unsettling, being scientifically as well as artistically relevant.

3. Script Generation using THEaiTRobot

We now move to describing and discussing the THEaiTRE project, both within the historical context and through the lens of today. In this section, we discuss the THEaiTRobot tool,[9] which we developed and used for generating the theatre play scripts. We note some of the limitations we had to face as well as various innovations which we introduced, and put them into context of the development in the field.

3.1. Turn-based Interactive Generation

THEaiTRobot was designed for generating theatre play scripts in an interactive human-in-the-loop fashion [29, 30]. First, the user (or *operator*) provides an input which sets the initial direction of the generation process: a setting of the scene, names of two characters, and their first lines (see Figure 1). Then the tool takes over, generates a chunk of text based on the user input (one line by default), and

[8] https://www.hendry.me/feelingmachine/
[9] https://github.com/ufal/theaitrobot

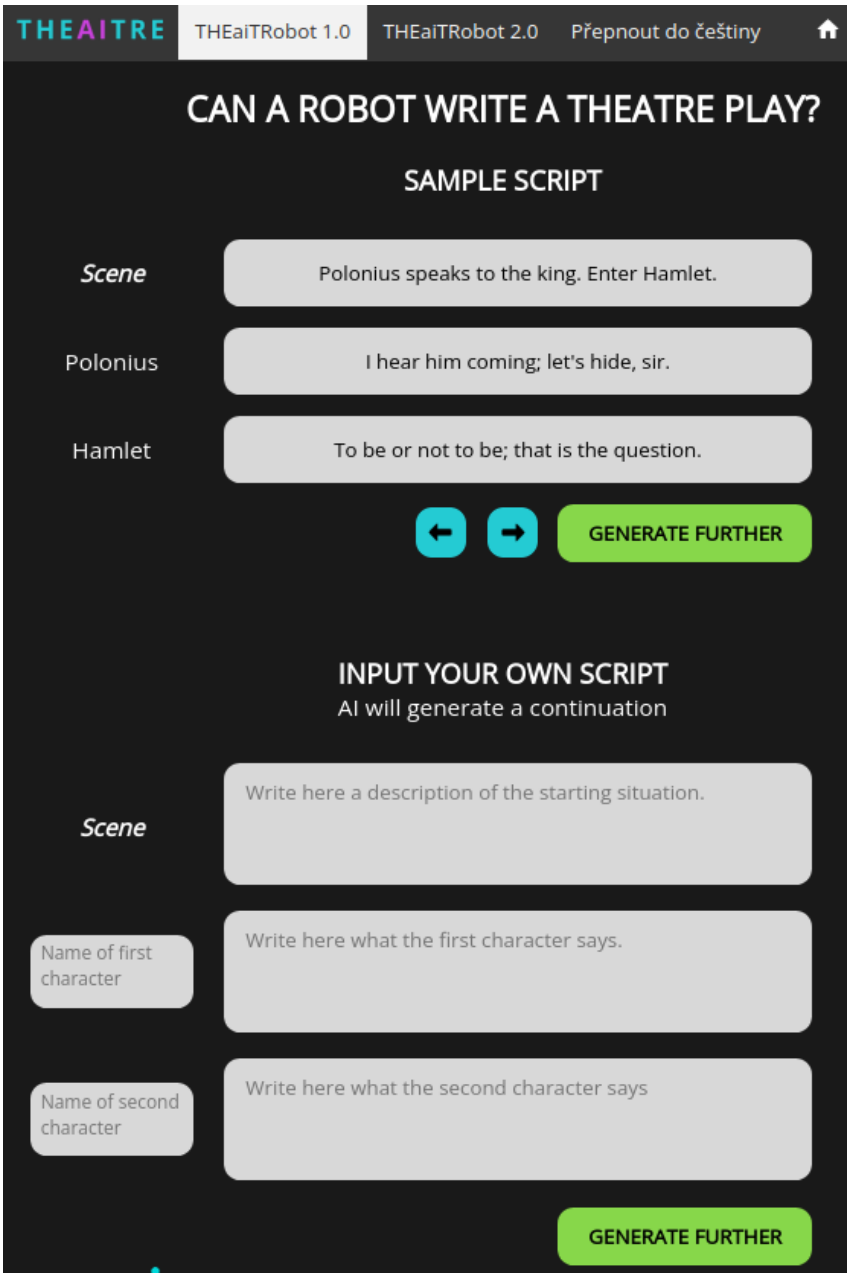

Figure 1: Screenshot of the THEaiTRobot input screen, with the option to either use a predefined script start, or to define one's own script start. (Self-made screenshot of the open-source application, licenced under the MIT licence.)

gives control back to the operator. The operator is then provided with two options of how to continue with the generation: either they like the script so far and choose to generate a following line, or they choose to discard the last generated line and generate another version (see Figure 2). It is thus an iterative process of automated generation guided by a human operator.

At the time of the project, this approach was new and revolutionary, as the standard way of using a LLM was not interactive or dialogical (as is usual today) but rather batch-based. The user would provide a prompt (a start of a piece of text), the LLM would generate a response (a completion of the text) ending with the end-of-document mark, and the interaction was over. Another notable project which diverged from the standard operation in a similar way was *AI Dungeon* [31], which was concurrent to our project. However, the dialogical chat-like interaction with LLMs only became usual with the introduction of ChatGPT in 2022.

The then-available models were also not yet instruction-tuned – this only came with the InstructGPT

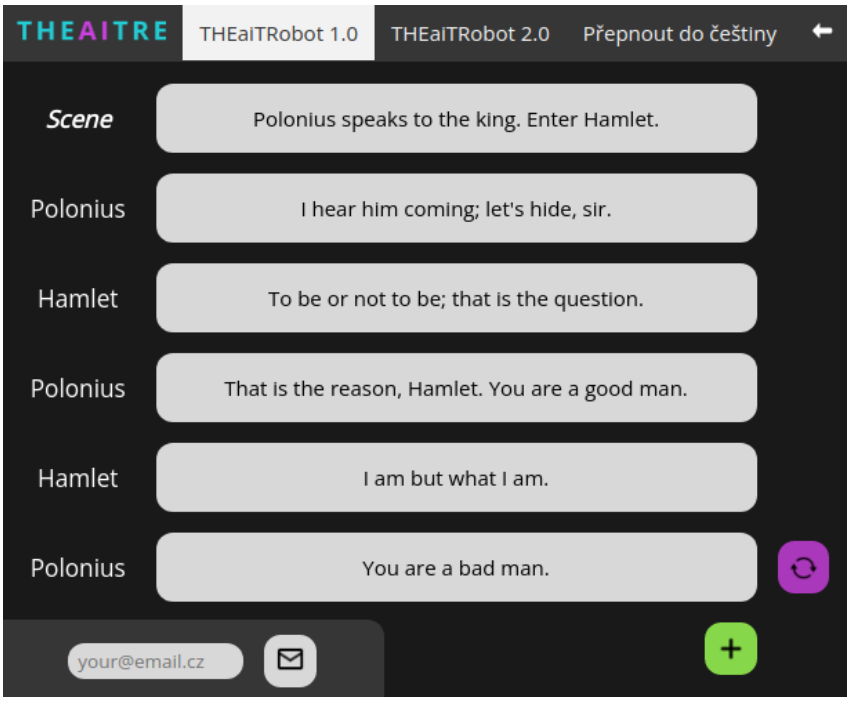

Figure 2: Screenshot of the THEaiTRobot generation screen, with several continuation lines already generated, and the buttons to either generate one further line or to discard and regenerate the last line. (Self-made screenshot of the open-source application, licenced under the MIT licence.)

approach [32] in 2022 – meaning that it was not possible to give the models instructions in natural language. The only mode was text-completion mode, and any task thus had to be formulated through a prompt in the form of a beginning of a textual document, which the model completed by generating a possible continuation of the document. We thus made sure to format the prompt like a beginning of a theatrical script, to encourage the model to continue by generating a continuation of the script (or at least a dialogue in general).

Nowadays, the main approach is inverse: instead of rephrasing your task to fit the training data, we reformat the training data to fit the envisioned future tasks. To make the LLMs understand instructions, data that can be understood as a task specification and completion are automatically reformatted as task instruction followed by task fulfillment; e.g. a news article that consists of a *title* and *text* can be reformatted into an instruction such as "Write a news article with the title *title*" followed by the *text*. The subsequent training process is the same, the underlying mode of operation is thus still text completion, but thanks to these additional examples in the training data, the models learn that a good text completion for a document starting with an instruction is the fulfillment of the instruction.

3.2. Mode of Human-Machine Interaction

During the THEaiTRE project, we experimented with various variants of putting the human into the loop of the script generation process. The first version of the application worked with the model always generating a continuation of only one line, but in five variants, and the operator then chose which continuation to use (or discarded all continuations and had another five continuations generated). To this the model again responded by generating five more continuation options after that selected line. This method of generation was therefore more interactive, but it did not work well in practice – on the one hand, it was too time-consuming to generate anything, and on the other hand, the theatrical experts operating the tool felt that they had too much control over the generation, and therefore the resulting script was more of their work rather than that of artificial intelligence. Generating only one candidate continuation at each step was eventually found to be both more efficient as well as not giving too much control to the operator.

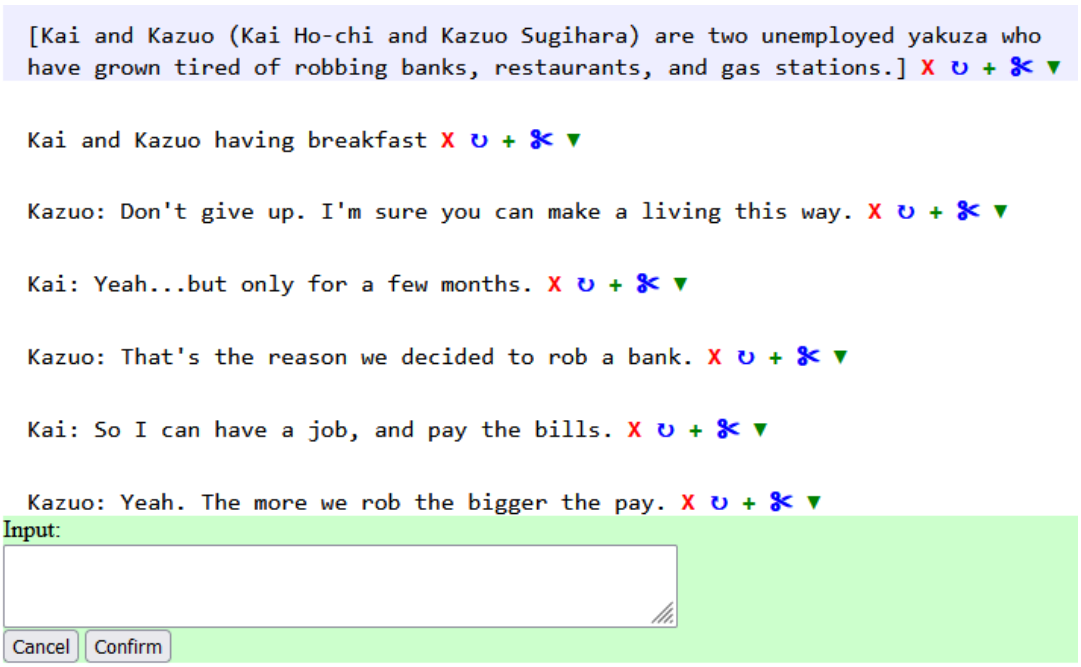

Figure 3: A screenshot from an internal advanced version of THEaiTRobot. The cross control (X) discards the given line and all subsequent lines and generates an alternative continuation. The round arrow control regenerates only the given line. The plus control (+) generates and inserts a line after the given line. The scissors control deletes the line. The triangle control (▾) allows for manually inserting human input. The cross and triangle controls lead to a regeneration of the continuation of the script, while the arrow, plus and scissors controls keep the continuation of the script unchanged. (Self-made screenshot of the open-source application, licenced under the MIT licence.)

This is well reflected in current LLM-based tools, which, even if unsure about the best output, typically do not offer multiple options but rather provide the user with only one generated response (with the option to discard and regenerate). Some tools sometimes offer two options for the user to choose from; but this is typically rather a form of A-B testing to have users compare between two versions of the tool.

We also found out that for an individual generated line, it is very difficult for the operator to decide whether it is good or not, while a longer section of the script is easier to judge and decide whether to accept or reject. Eventually, the tool was modified to always generate 10 lines in one step, as a compromise between the speed of generation and the degree of control over the generation process. Tuning optimal output length to user expectations may still be difficult today, as users may prefer to receive a shorter output at first to see whether the generation is going in the right direction and to have continual control over the process – instead of generating several pages of text at once which they may then need to immediately discard if the output does not fulfill their desires [33].

In a later, more advanced version of THEaiTRobot, we also introduced more controls for the operator, allowing not only to regenerate the last generated line, but also to regenerate or delete any of the previously generated lines, or to manually insert a line anywhere into the script (see Figure 3). This would of course not influence the other already generated lines, but would be taken into account by the model when generating new lines. We note that such control options are even nowadays not usual in the mainstream LLM-based tools which typically do not allow the user to change the previous context in any way, although our experience showed them to be highly useful for precisely targeted edits of a longer piece of text. The current usual mode of operation is to issue textual editing instructions to the model, which then reproduces its previous output with changes according to the instructions; however, the user does not have full control over the possibility that the model also performs additional changes which the user did not request and might not want or even notice.

3.3. Training Data

THEaiTRobot was based on the large pretrained generative neural language model called GPT-2 [18], a predecessor of the currently widely known and popular family of GPT models and tools, such as ChatGPT. GPT-2 had been trained on 40GB of English texts and was shown to be very strong in generating new high-quality texts based on even a short initial prompt. However, as most of the training texts were not theatrical scripts but rather news texts and other typically short texts with a simple structure, the main research challenge in our work was to adapt the GPT-2 model to generating theatre scripts, which are rather long and complexly structured.

We mainly used movie scripts as a proxy, as these were available in much larger quantities for training the models; however, this made the generated theatre scripts somewhat too movie-like, which had to be dealt with by the theatrical team.

Although the large language models have improved dramatically since the project, their limitations imposed by training data are still today hard to overcome. A language model can produce texts similar to its training data with high quality, but the more we diverge from the kinds of texts abundant in its training data, the worse the results get. Current large language models can abstract from the data to an impressive extent, producing good novel outputs very different from its training data, but (at least under the current machine learning paradigm) this will always have its limits. Examples of these limits are the well-known hallucinations of the models, most prominent when the models generate text the domain/topic of which is not well represented in their training data [34], or the significant downgrade of quality of the outputs generated in low-resource or no-resource languages (languages that form only a small part of the training data, or are not present at all in the training data) [35].

The language limitation was particularly stark with GPT-2, as it was only trained on English data and incapable of working with texts in other languages. For this reason, the scripts were generated in English and then automatically translated to Czech, as the goal of the project was to produce a theatre play in Czech language. The automated translation obviously introduced an additional amount of noisiness of the text, which had to be remedied to some extent in manual post-editing. Moreover, the English-only model struggled with Czech names of characters; usually, it would assume they are Indian or Russian names, which would shift the play into a setting which we did not want (e.g. naming a character Maria would often lead to the play taking place in Moscow). Therefore, the characters in the play do not have names but are referred to simply as Robot, Boy, Psychologist, etc.

With current massively multilingual language models, the language limitation is greatly diminished, as the models are capable of operating in dozens of languages with satisfactory quality. However, the problem of language limitation is more hidden than resolved. Only a small fraction of the world's languages are covered, and language performance greatly depends on the size and quality of available training data; this typically means that the models perform much better in English than in any other language. Thus, while within the THEaiTRE project, the language limitations were completely overt, they are not so easy to clearly view at first sight nowadays.

3.4. Limited Context window

A script of a theatre play is a long and complex textual document. However, the context window of GPT-2 was only 1,024 tokens for both input and output combined (so approximately half a page of text). Once the generation ran out of the context window, the beginning of the script got clipped and the generation started losing context. The generated text was thus locally coherent, but did not hold well together in the longer distance, gradually drifting away from the original theme, becoming contradictory, etc.

For the first play, we therefore generated the script for each scene independently, as there was no simple way of maintaining the context of the whole play within such a limited window. Several scenes were generated independently, bound together only through the person of the operator, who specified the settings and characters in such a way that some basic consistency is maintained: In all the scenes, there are two characters, where one of them is the Robot, and the other is a different person each time

who the Robot encounters and interacts with in a specific setting (akin to The Little Prince [36]) – e.g. a schoolboy who seeks life advice, or a sex worker who tries to seduce the Robot. The scenes are thus bound together through the central character and the common format, but this consistency is not maintained through the tool but through its human operator. The operator also tried to guide the generation process in a way that the scenes can be then meaningfully put together, but in practice this was often impossible, as they had only very limited amount of control over the generation process. For the first play, scripts of 15 scenes were generated. In the staging process, 8 of them were selected and juxtaposed to form the script of the whole play.

Even then, a whole scene script would not fit into the context window of the language model. Therefore, we applied a simple summarization approach, always keeping the beginning and end of the script but removing some of the lines in the middle based on automated importance ranking.

Since GPT-2, the size of the context window is an always expanding parameter, with current models supporting contexts of millions of tokens or more (dozens or even hundreds of pages of text) [37]. However, for large tasks, the context window can still be insufficient, which still necessitates some clipping and/or summarization to be applied; presumably more sophisticated than the approach we had invented, but the exact mechanism is not disclosed for the commercial models [38, 39].

For the second play, we introduced a hierarchical approach to bypass the limited context window [40]. It was a two-step approach, consisting of synopsis generation and script generation. In the first step, a high-level synopsis of the whole play was generated by a model trained on plot summaries, and cut up into individual sentences. The subsequent script generation step was similar to the script generation used for the first play, with the additional operation of injecting sentences from the synopsis to condition the generation by the synopsis whenever the operator decided that the play should move on to the next point in the synopsis. Still, the generation was not conditioned by the "future" lines of the synopsis (not yet shifted into the script) and thus would continually diverge from it, making the process reportedly rather frustrating and leading to more discontinuities in the script than when the previous approach was used. Therefore, the hierarchical approach was eventually not used for generating the script of the second play.

Although the context window is now not limiting for documents of the size of a theatre play, in practice, it has been found that models tend to struggle with long unstructered contexts, as well as with making complex decisions within a single step of generating a single token. Therefore, it has become common that multi-step approaches are used, such as chain of thought or reasoning [41, 42].

4. Staging and Reception

Once the script of the play was generated and finalized, it was provided to the theatre company.

The play AI: When a Robot Writes a Play was rehearsed and staged in a classical way, with a full stage production involving 6 actors, a scenographer, costumes, music, etc. All of the staging was done manually, with no involvement of artificial intelligence [43]. The play lasts approximately 60 minutes and consists of 8 scenes. The play premiered in 2021 and stayed on the repertoire of the theatre for two years, with performances approximately once a month.[10]

The Permeation play was only performed twice in the form of stage reading, as we did not find it to add too much to what we already achieved with the first play.

4.1. Reception by Co-creators

When working with the initial versions of the application, the theatrical team was rather disappointed by the low quality of the outputs and planned to stage the play in a less expensive form of one-off stage reading without decorations and costumes and without further performances. However, through the gradual development of the application and also through the gradual closer familiarization of the theatrical team with the application and a better understanding of its functioning and its strengths and

[10] https://www.svandovodivadlo.cz/inscenace/673/ai-kdyz-robot-pise-hru/3445

weaknesses, it was possible to achieve a state where the theatrical team created, according to their own opinion, a sufficiently high-quality script using the application.

In retrospect, we now believe this to be an early example of a phenomenon that is being repeatedly attested across various fields. Experience shows that, especially among artistic and humanities experts with low/no technical background, the first user experience with generative AI systems is often negative, as they use the systems in a naive way, receive unsatisfactory outputs, and, without further attempts at refining the result through dialogical engagement with the system, immediately dismiss the system (and potentially the whole field) as low-quality and useless. Artists and creators also often report having low agency in the process, not feeling to be in control over the result, and thus not having a meaningful creative contribution and not feeling as a (co-)author of the output. However, as users gain more experience with the AI tools and more technological understanding of the systems and the process, the attitude typically shifts towards being more positive. In particular, technologically experienced creators who are using the AI systems in an advanced way, and especially if they also participate in the design of the system, usually report that they are in control of the process, that they are meaningfully collaborating with the tool, and feel as authors or co-authors of the resulting works [44, 45, 46, 47].

Nevertheless, the theatre professionals – the operator of THEaiTRobot, the director, and the actors – viewed the collaboration with the tool as well as working with the generated script mostly negatively. They noted that the process of creating the script was more laborous and tedious than creating a script manually, while achieving a much worse result; although they also noted that such a tool might be useful for some minor usage, e.g. to overcome writers block by suggesting various alternatives and potential developments. They suggested that while in our project, the ratio of generated versus human-provided content was 90:10, they could imagine that a reverse ratio of using 10% generated content might be useful and interesting. This is quite in line with the current situation in creative fields, where professional creators typically do not generate most of the content of their artwork automatically, but rather use the AI tools as support assistants that only provide a minor amount of the content of the work (if any) and perform various auxiliary tasks.

The rehearsing and performing of the generated script proved to be particularly unpleasant for the director and the actors. One of the concrete problems were the logical gaps and discontinuities in the script, where the team could not fully make sense of the script, the actors could not understand their characters and feel into them, and even with a lot of interpretative hard work, there were always many gaps remaining where the actors simply needed to grit their teeth and leap over. On the other hand, for many of the superficially illogical and nonsensical places, the team eventually devised a meaningful interpretation ('around' the situation rather than 'of' the situation) which they managed to convey through ingenious acting (as they were not allowed to modify the script). This shows that, while often unpleasant, generated content can actually push creative individuals into novel and interesting outcomes. This is often used in other projects where AI tools are used as partners in improvised theatre.

As for the more vague claims, many members of the team complained that the script does not have a 'soul' (note that this was expressed even by several atheists). The team complained that working with the script does not provide them with any positive feelings and energy, as is usual with human-written scripts, but rather drains them: it does not give them anything, and they need to give a lot to it. They also expressed that with a human-written script, there positively is an author's intention behind the script, however obscure, which they may try looking for, and even if they do not arrive at the original intention, they perceive traces of the intention in the script and enjoy the process. On the other hand, with the machine-generated script, they *know* there is no intention and there is nothing behind to try to look for. If there is a part of the script that does not seem to make sense, it is not worth trying to imagine what the author might have had in mind so that the script would actually make sense; it is *clear* that there is no hidden sense behind and that to make sense out of it, one cannot find the sense behind the script, but rather has to invent the sense and add it to the play. In general, the team seemed to be missing some transcendental aspects of the script of there being something more beyond the script. It must be noted though that all the participants were aware that the script is AI-generated, and we did thus probably observe an early instance of the "AI stigma" or implicit bias against AI. At the same time, the quality of the script was indeed not great, and the GPT-2 model was probably too simple for

there to be anything interesting behind what was overtly present in the script; but this may not be so straightforward with current powerful models (and even more so with the reasoning models).

These aspects of generated content are widely and sometimes even wildly debated nowadays, much more than at the time of the THEaiTRE project [48, 49, 50]. Questions such as whether LLMs can truly communicate, whether they can have real intent, whether we can ascribe grounded meaning to their outputs, whether they really understand language, etc., are still largely open, with the debates hindered by the lack of agreement on the meaning and/or definition of many of the concepts that are being discussed. The discussion is further polluted by the latent conviction of some of the debaters that AI systems are inherently different from humans and thus any conclusion equating some aspects of humans and machines is a priori inacceptable; they are thus not debating to find the truth, but rather to convince others of their already presupposed belief. However, it is clear that many feel strongly about these aspects of AI, having strong relevance towards value of artwork, interpretation of the artifacts, and in general being relevant to the perception and enjoyment of the works.

4.2. Reception by Audiences

Except for the online premiere, the play was performed in front of live audiences. We noted that the spectators were typically engaged and attentive, enjoyed the performance, did not get too bored to leave the theatre during the play and often laughed at various points of the performance. After the performance, they were often surprised that the 60 minutes had already passed; this exceeded our expectations, as we had been dubious whether the play based on a generated script can hold the attention of the spectators for so long. However, it is clear that a lot was done by the staging of the play, adding some meaning to the script through acting and leaping over some of the nonsensical gaps in the script. The rating of the play by spectators on a major Czech drama rating website is 61%,[11] which ranks it into the bottom quartile of the rated plays; however, there are many human-made plays on the website that are ranked worse than our play (including some plays by the same theatre).

It is important to note that the play was produced before the ChatGPT-induced AI boom, and thus was interesting to the audiences by having an AI-generated script, was actively promoted as an AI-generated play, and the involvement of AI was the main reason why spectators decided to see the play. People were generally suspicious of the capabilities of AI systems to produce a play (and rightfully so, as the quality was still low given the small simple models); but still there was also great curiosity. This is of course very different from the current situation, as the AI enthusiasm has been gradually shifting to boredom by the omnipresent AI, eventually moving towards open AI hate.

A number of professional reviews of the play were also published by experienced dramatical reviewers;[12] they were somewhat mixed and prevalently negative, but all of them also contained some positive surprisal and appreciation. It is obvious that the script is not entirely logical, much is inevitably left to the spectator's interpretation, and in some cases it is probably very difficult or impossible to find a satisfactory interpretation. However, the fact that this result of the project can actually be viewed as a full-fledged play and thus evaluated, is considered to be a great success by our team. The reviews generally note the imperfections of the AI author, but in general, the quality of the play seems to be in some sense comparable to the quality of plays by a below-average or bad human author.

Some reviewers also noted that the play contains many common human biases and stereotypes [8]; it is now well known that large language models often tend to be e.g. racist or sexist [51]. The general cause is the fact that the underlying training datasets (articles downloaded from the internet with light

[11] https://www.i-divadlo.cz/divadlo/svandovo-divadlo/ai-kdyz-robot-pise-hru

[12] https://www.itechpost.com/articles/104880/20210301/elon-musks-openai-system-creates-first-ai-written-theater-script.htm,
https://www.britishtheatreguide.info/reviews/ai-when-a-robo-19769,
https://www.ayoungertheatre.com/review-ai-when-a-robot-writes-a-play-czech-centre-london/,
https://www.lidovky.cz/kultura/recenze-robot-pise-a-la-beckett-kde-vsechno-tohle-umela-inteligence-pochytila.A210301_112539_ln_kultura_jto?,
https://www.idnes.cz/kultura/divadlo/ai-kdyz-robot-pise-hru-recenze-svandovo-divadlo.A210301_105549_divadlo_ts?,
https://www.i-divadlo.cz/blogy/svatosova/robot-a-jeho-doba

or no filtering) themselves contain such biases and other problematic content, and the language model merely truthfully reproduces those. As we intended to give a truthful image of the current state of the technology, we intentionally kept some of this problematic content in the final script. However, we made sure to point to these issues and explain them in discussions after the performances and at other public events related to the project.

Since the release of ChatGPT, it has become standard that the language models are not simply trained on unfiltered texts from the whole internet, but various measures are employed to try to make the generated outputs safer and less problematic [52], such as data filtering [53], reinforcement learning with human feedback [54], and debiasing [55]. Therefore, most currently available LLMs produce significantly less problematic outputs than the GPT-2 model we had used; although the issue of potentially producing problematic content is still present even with the top LLMs and may never be fully resolvable.

An integral part of the performance was an expert discussion that followed each showing of the play. The discussion always involved experts from the THEaiTRE project team, who introduced the project, discussed with the audiences and answered their questions. This way, we made sure to fulfill one of the important goals of the project, which was to bring closer and explain artificial intelligence to people. The experience from the discussions showed that this was largely successful, the discussions tended to be stimulating and factual, and an increasing level of understanding was observed in the participants during the discussion. It is interesting that some participants unequivocally rated the discussion as the better part of the evening than the play itself. We thus concluded that the play only has its complete meaning and fulfills its full potential when accompanied by the professional discussion.

In recent years, we have repeatedly seen that people show disinterest into AI-generated works that stand on their own, with no obvious creative human involvement in the process; people require "the human touch" somewhere in the process to appreciate the result [56, 57]. This may be in the initial idea, in the collaborative creation process, and/or in a human performing the generated work; but it is vital for the spectators/perceptors that it is there and that they know it is there and where it is.

Actually, it has been repeatedly shown that the belief that the work is created or co-created by a human, rather than fully generated, is more important than whether this is true. People tend to rate generated works better if they are marked as human-made, and vice-versa; and may actually rate AI-generated works higher than human-created works in blinded studies [58, 59, 60]. This is becoming to be known as the "AI stigma": people tend to be biased against the quality of works if they believe that they are partially or fully AI-generated.

We are also observing an increased focus on authenticity [61], going as far as to creators recording their creation process on video and publishing it on social media to share the process with the perceptors of their art and to prove that the artwork is authentic. In this context, although our goal in the project was to minimize human involvement in the process, by having failed to do so completely (partially in the script and especially in the play production) we may have inadvertently succeeded.

5. Findings and Conclusion

In this paper, we looked back onto the THEaiTRE project, in which we produced a theatre play with an AI-generated script. We described the setting of the project within its historical context, explaining how the project was carried out, and viewing it through the lens of the AI boom that ensued after the end of the project. We discussed how the view of many of the problems, concepts and methods developed over time, identifying cases where development followed in similar footsteps as our work, as well as cases where the field went in a different direction. We also observed how the views on AI and its usage in art production have been evolving in the past few years, together with various problems that are still not fully resolved. We highlight some of our main findings and observations with broader relevance.

From curiosity to hatred At the time of conception of the project, AI was far from being mainstream, the main showcase of computational linguistics was machine translation (which was not typically thought of as an example of AI), and the popular conceptions of artificial intelligence were more

based around rule-based processing and physical robots. A project putting AI into theatre was thus viewed as a curiosity, something improbable or possibly even impossible; mostly nobody expected the AI boom which was already imminent. Since then, we have observed how AI has risen and proliferated into many fields of human activity, with the sentiment going from curiosity and fascination (or ignorance and dismissal), over acceptance, then boredom, and eventually hatred (among the general population) and stigma (especially in the art world). Interestingly, we observed the same phases within the project itself in a fast pace, reaching the hatred phase among many of the actors already in 2021.

Scripted vs improvised In our project, we aimed at generating a script for a classical theatre play beforehand, while other projects focused on interactive employment of AI as a collaborative partner in improvisation on stage. While typical practical usage of AI tools nowadays is to prepare rather classical human-like texts for further circulation (akin to scripted theatre), this has only very limited value and success in theatrical practice (as well as in other arts). Artists are rather interested in the inhumanity of the AI, the glitch, etc., employing AI rather in experimental than classical settings, shifting the focus on the AI itself, e.g. by engaging with an AI system live on stage within the setting of improvised theatre.

Dialogical interaction and human-in-the-loop In our project, we were among the pioneers of interacting with a language model within the form of a dialogue. This is a very natural choice for the theatrical setting, although rather unnatural for a language model trained primarily for text completion. Interestingly, this eventually turned out to be one of the major appeals facilitating human interaction with language models and making it considerably more natural for humans, to the extent that dialogical interaction with AI chatbots has become the dominant paradigm. On the other hand, AI systems are prevalently not fully replacing humans, but the dominant and most successful approaches are rather of the human-in-the-loop nature: as we already observed in THEaiTRE, it is easy to minimize the human involvement to some extent, but there typically is a tipping point where further minimization of human involvement in the process inevitably leads to strong deterioration in quality.

Anthropomorphization of AI vs human exceptionalism Somewhat unsurprisingly, the dialogical nature of interaction with AI systems leads to (or at least strengthens) the Eliza effect [62]. Through millenia, people as a social species evolved to assign self-experienced human traits (such as consciousness or intentionality) to other entities that show human-like behaviour. In the face of conversational AI that exhibits intelligent behaviour, this inadvertently but inevitably leads to anthropomorphization of AI. In the scientific community, this is predominantly viewed as a negative effect, as it leads people to assume traits which the systems do not have. However, within THEaiTRE project, we also observed that the GPT-2 model was much more human-like than people expected; instead of a seemingly cold and rational machine, they encountered a system simulating various human imperfections, such as racism or sexism. While over-anthropomorphization of AI is clearly harmful, the opposing anti-anthropomorphization and human exceptionalism are also not unproblematic. After all, the models are trained on data produced by humans, and thus some human traits may be observed within the system, either simulated or even emergent. In the face of an 'alien intelligence', which language models can be thought to be an example of, some amount of anthropomorphization (or 'demonomorphization' [63]) may be our best attempt at conceptualizing and interacting with the entity. Strong human-centric claims of the inherent impossibility of AI systems to exhibit e.g. intentional behaviour may not be well grounded and need further discussion and research. Importantly, this has direct implications to various other human qualities that are important for the value of art, such as creativity, authorship, or authenticity.

Human agency and authorship Already in THEaiTRE, through experimenting with various modes of human-machine interaction, we observed a variance in the human operator's feeling of their

own agency, of being in control. When generating line by line, with 5 options generated for each line, the operators felt to be 'too much' in control, to the extent that they would claim that they, not the AI system, are actual authors of the text. With a larger chunk being generated at once with little other control, the humans felt as mere operators, with the AI doing the bulk of the writing, and thus being the main author. The complaint of lacking sufficient control over the creative process, and thus not meaningfully fulfilling the creative authorship role, resonates within the artistic field; the voices are especially strong with technically unskilled artists using the AI systems in a simplistic naive way ('single click generation') and thus truly having very little control over the process. However, artists with more AI experiences, willing to embrace the tools, often feel a higher level of agency and control, especially if they are actively involved even in the process of creating/training the system. Relatedly, technically more experienced creators are often to produce better outputs with which they are more satisfied and feel more as the authors or co-authors.

Artistic research Finally, while the initial AI-theatre projects were rather applied and exploratory, they seem to have laid ground for more sophisticated interdisciplinary endeavours of artistic research. Within the project itself, we did not arrive at many scientifically important findings; some interesting insights are only revealed through time. However, some of the current projects already successfully use theatrical methods to perform research on AI systems (and vice versa), such as using theatrical skills of character embodiment to better understand AI systems. We believe this to be probably the most fascinating outcome of all of what has happened within the AI-theatre interaction.

While the THEaiTRE project itself was probably more of a dead-end from today's perspective, we believe that many of its aspects and especially some of our findings and interpretations of the project are interesting and relevant for the current development and discussions of the relation between artificial intelligence and art.

Acknowledgments

The research was supported by Czech Science Foundation Grant No. 25-14501L *AI: Authorship and Interpretation*. The original project TL03000348 *THEaiTRE: Umělá inteligence autorem divadelní hry* was co-financed with the state support of Technological Agency of the Czech Republic within the ETA 3 Programme (2020-2022). The work described herein has also been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

Generative AI Statement

During the preparation of this work, the author used ChatGPT for the tasks of: Grammar and spelling check (identification of errors and suggestion of corrections, used for manually improving the text), Peer review simulation (analysis of the paper and suggestion of potential reviewer comments, used to manually improve the first version of the paper), Generate literature review (search and summarization of relevant publications, which were then manually reviewed and referred to in the text if appropriate). All of the text was written manually, with the outputs of ChatGPT being used only as suggestions which the author reviewed and selectively manually incorporated into the text. The author takes full responsibility for the publication's content.

References

[1] R. Rosa, O. Dušek, T. Kocmi, D. Mareček, T. Musil, P. Schmidtová, D. Jurko, O. Bojar, D. Hrbek, D. Košťák, M. Kinská, J. Doležal, K. Vosecká, THEaiTRE: Artificial intelligence to write a theatre

play, in: A. Jorge, R. Campos, A. Jatowt, A. Aizawa (Eds.), Proceedings of AI4Narratives — Workshop on Artificial Intelligence for Narratives, volume 2794 of *CEUR Workshop Proceedings*, RWTH Aachen University, RWTH Aachen University, Aachen, Germany, 2020, pp. 9–13.
[2] P. Schmidtová, R. Rosa, D. Košťák, T. Studeník, D. Hrbek, T. Musil, J. Doležal, O. Dušek, D. Mareček, K. Vosecká, M. Nováková, P. Žabka, A. Zakhtarenko, D. Jurko, M. Kinská, T. Kocmi, O. Bojar, THEaiTRE: Generating Theatre Play Scripts using Artificial Intelligence, Institute of Formal and Applied Linguistics, Prague, Czechia, 2022. URL: https://ufal.mff.cuni.cz/books/2022-schmidtova.
[3] K. Čapek, R.U.R. (Rossum's Universal Robots), Aventinum, Ot. Štorch-Marien, Praha, 1920.
[4] R. Rosa, P. Schmidtová, A. Zakhtarenko, O. Dušek, T. Musil, D. Mareček, S. O. ul Islam, M. Nováková, K. Vosecká, D. Hrbek, D. Košťák, THEaiTRobot: An interactive tool for generating theatre play scripts, in: Proceedings of the 15th International Conference on Natural Language Generation: System Demonstrations, Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022, pp. 10–13.
[5] R. Rosa, T. Musil, O. Dušek, D. Jurko, P. Schmidtová, D. Mareček, O. Bojar, T. Kocmi, D. Hrbek, D. Košťák, et al., When a robot writes a play: Automatically generating a theatre play script, in: ALIFE 2021: The 2021 Conference on Artificial Life, MIT Press, 2021, pp. 1–3. doi:https://doi.org/10.1162/isal_a_00372.
[6] THEaiTRobot 1.0, D. Košťák, D. Hrbek, R. Rosa, O. Dušek, AI: When a Robot Writes a Play, Technical Report ÚFAL TR-2021-67, ÚFAL MFF UK, Praha, Czechia, 2021. URL: https://ufal.mff.cuni.cz/techrep/tr67.pdf.
[7] THEaiTRobot 2.0, J. Doležal, K. Vosecká, T. Musil, D. Mareček, R. Rosa, Permeation, Technical Report ÚFAL TR-2022-71, ÚFAL MFF UK, Praha, Czechia, 2022. URL: https://ufal.mff.cuni.cz/techrep/tr71.pdf.
[8] A. Akbar, On the scene, like a sex-obsessed machine: when a robot writes a play, The Guardian (2021). URL: https://www.theguardian.com/stage/2021/mar/01/on-the-scene-like-a-sex-obsessed-machine-when-a-robot-writes-a-play-ai.
[9] M. Colonescu, Can an AI be taught to write a play?, Sustained independent project, Central School of Speech and Drama, University of London, 2022.
[10] I. Van Heerden, Ç. Duman, A. Bas, Performing the post-anthropocene: AI: when a robot writes a play, TDR 67 (2023) 104–120.
[11] S. Efimova, KI-textgenerierung und autor: innenschaft am beispiel von AI: When a robot writes a play (2021), in: Autor: innenschaft und/als Arbeit, Brill Fink, 2024, pp. 117–130.
[12] S. Ç. İleri, The evolution of the theatre play, human to AI: The case of AI: When a robot writes a play (2021), ETKİ: Journal of Literature, Theatre and Culture Studies 5 (2025) 29–40.
[13] S. Elias, B. S. Alshammari, K. N. Alfraidi, K. M. Karam, Rethinking literary creativity in the digital age: a comparative study of human versus AI playwriting, Humanities and Social Sciences Communications 12 (2025) 1–11.
[14] H. Nasir, Artificial intelligence as a tool for reimagining experience: A study of AI: When a robot writes a play, Central Asian Journal of Literature, Philosophy and Culture 7 (2026) 28–36.
[15] M. Sundermeyer, H. Ney, R. Schlüter, From feedforward to recurrent lstm neural networks for language modeling, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2015) 517–529.
[16] S. Colton, M. T. Llano, R. Hepworth, J. Charnley, C. V. Gale, A. Baron, F. Pachet, P. Roy, P. Gervás, N. Collins, B. Sturm, T. Weyde, D. Wolff, J. R. Lloyd, The Beyond the Fence musical and Computer Says Show documentary, in: Proceedings of the Seventh International Conference on Computational Creativity, 2016, pp. 311–320.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 1–11. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsu-

pervised Multitask Learners, Technical Report, OpenAI, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[19] K. W. Mathewson, P. Mirowski, Improvised theatre alongside artificial intelligences, in: Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference, 2017, pp. 66–72.
[20] K. W. Mathewson, P. Mirowski, Improvised comedy as a turing test, CoRR abs/1711.08819 (2017). URL: http://arxiv.org/abs/1711.08819. arXiv:1711.08819.
[21] K. Mathewson, P. Mirowski, Improbotics: Exploring the imitation game using machine intelligence in improvised theatre, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 14, 2018, pp. 59–66.
[22] P. Mirowski, K. Mathewson, B. Branch, The theater stage as laboratory: review of real-time comedy LLM systems for live performance, in: Proceedings of the 1st Workshop on Computational Humor (CHum), 2025, pp. 88–95.
[23] P. Mirowski, K. Mathewson, B. Branch, T. Winters, B. Verhoeven, J. Elfving, Rosetta code: Improv in any language, in: Proceedings of the 11th International Conference on Computational Creativity, Association for Computational Creativity, 2020, pp. 115–122.
[24] B. Branch, P. Mirowski, K. W. Mathewson, Collaborative storytelling with human actors and AI narrators, in: Proceedings of the 12th International Conference on Computational Creativity, Association for Computational Creativity, 2021, pp. 97–101.
[25] G. Lösel, Theatre dialogues with machines, Interdisciplinary Science Reviews 49 (2024) 291–304. URL: https://doi.org/10.1177/03080188241257882. doi:10.1177/03080188241257882. arXiv:https://doi.org/10.1177/03080188241257882.
[26] D. Mareček, J. Libovický, T. Musil, R. Rosa, T. Limisiewicz, Hidden in the Layers: Interpretation of Neural Networks for Natural Language Processing, volume 20 of *Studies in Computational and Theoretical Linguistics*, Institute of Formal and Applied Linguistics, Prague, Czechia, 2020.
[27] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.
[28] M. F. Hendry, N. Kottmann, M. Fröhlich, F. Bruggisser, M. Quandt, S. Speziali, V. Huber, C. Salter, Are you talking to me? a case study in emotional human-machine interaction, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 19, 2023, pp. 417–424.
[29] R. Rosa, T. Musil, O. Dušek, D. Jurko, P. Schmidtová, D. Mareček, O. Bojar, T. Kocmi, D. Hrbek, D. Košťák, M. Kinská, M. Nováková, J. Doležal, K. Vosecká, T. Studeník, P. Žabka, THEaiTRE 1.0: Interactive generation of theatre play scripts, in: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (Eds.), Proceedings of the Text2Story'21 Workshop, volume 2860 of *CEUR Workshop Proceedings*, RWTH Aachen University, Aachen, Germany, 2021, pp. 71–76. URL: http://ceur-ws.org/Vol-2860/paper9.pdf.
[30] R. Rosa, P. Schmidtová, O. Dušek, T. Musil, D. Mareček, S. O. ul Islam, M. Nováková, K. Vosecká, J. Doležal, GPT-2-based human-in-the-loop theatre play script generation, in: E. Clark, F. Brahman, M. Iyyer (Eds.), Proceedings of the 4th Workshop of Narrative Understanding, Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022, pp. 29–37.
[31] M. Hua, R. Raley, Playing with unicorns: AI dungeon and citizen NLP., DHQ: Digital Humanities Quarterly 14 (2020).
[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.
[33] M. X. Liu, F. Liu, A. J. Fiannaca, T. Koo, L. Dixon, M. Terry, C. J. Cai, We need structured output: Towards user-centered constraints on large language model output, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–9.
[34] A. Ravichander, S. Ghela, D. Wadden, Y. Choi, Halogen: Fantastic LLM hallucinations and where to find them, in: Proceedings of the 63rd Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), 2025, pp. 1402–1425.
[35] Z. Li, Y. Shi, Z. Liu, F. Yang, A. Payani, N. Liu, M. Du, Language ranker: A metric for quantifying LLM performance across high and low-resource languages, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 28186–28194.
[36] A. d. Saint-Exupéry, The little prince, New York (1943).
[37] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, M. Yang, LongRoPE: extending LLM context window beyond 2 million tokens, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 11091–11104.
[38] Y. Li, B. Dong, F. Guerin, C. Lin, Compressing context to enhance inference efficiency of large language models, in: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 6342–6353.
[39] M. Lu, W. Sun, W. Du, Z. Ling, X. Yao, K. Liu, J. Chen, Scaling LLM multi-turn RL with end-to-end summarization-based context management, arXiv preprint arXiv:2510.06727 (2025).
[40] P. Schmidtová, D. Javorský, C. Mikláš, T. Musil, R. Rosa, O. Dušek, DialogueScript: Using dialogue agents to produce a script, 2022. URL: https://arxiv.org/abs/2206.08425. arXiv:2206.08425.
[41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[42] Z. Ke, F. Jiao, Y. Ming, X.-P. Nguyen, A. Xu, D. X. Long, M. Li, C. Qin, P. Wang, S. Savarese, et al., A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems, arXiv preprint arXiv:2504.09037 (2025).
[43] T. Musil, K. Vosecká, R. Rosa, Performing AI-Generated Theater Plays, first ed., CRC Press, Boca Raton, FL, USA, 2023, pp. 390–400.
[44] L. Sun, Y. Liu, G. Joseph, Z. Yu, H. Zhu, S. P. Dow, Comparing experts and novices for AI data work: Insights on allocating human intelligence to design a conversational agent, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 10 (2022). URL: https://par.nsf.gov/biblio/10374260. doi:10.1609/hcomp.v10i1.21999.
[45] V. van der Burg, G. de Boer, A. A. Akdag Salah, S. Chandrasegaran, P. Lloyd, Objective portrait: A practice-based inquiry to explore AI as a reflective design partner, in: Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 387–400. URL: https://doi.org/10.1145/3563657.3595974. doi:10.1145/3563657.3595974.
[46] E. Eigner, T. Händler, Determinants of LLM-assisted decision-making, arXiv preprint arXiv:2402.17385 (2024).
[47] Y. Feng, Y. Sun, Consumer empowerment through generative artificial intelligence: enhancing brand narrative with collaboration, creation, and communication, International Journal of Advertising 45 (2026) 229–261. doi:10.1080/02650487.2025.2547545.
[48] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: https://aclanthology.org/2020.acl-main.463. doi:10.18653/v1/2020.acl-main.463.
[49] A. Søgaard, Grounding the vector space of an octopus: Word meaning from raw text, Minds and Machines 33 (2023) 33–54.
[50] M. Coeckelbergh, D. Gunkel, Communicative AI: A critical introduction to large Language models, Polity Press, 2025.
[51] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, RealToxicityPrompts: Evaluating neural toxic degeneration in language models, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3356–3369.
[52] K. Wang, G. Zhang, Z. Zhou, J. Wu, M. Yu, S. Zhao, C. Yin, J. Fu, Y. Yan, H. Luo, L. Lin, Z. Xu, H. Lu, X. Cao, X. Zhou, W. Jin, F. Meng, S. Xu, J. Mao, Y. Wang, H. Wu, M. Wang, F. Zhang, J. Fang, W. Qu, Y. Liu, C. Liu, Y. Zhang, Q. Li, C. Guo, Y. Qin, Z. Fan, K. Wang, Y. Ding, D. Hong, J. Ji, Y. Lai, Z. Yu, X. Li, Y. Jiang, Y. Li, X. Deng, J. Wu, D. Wang, Y. Huang, Y. Guo, J. tse Huang,

Q. Wang, X. Jin, W. Wang, D. Liu, Y. Yue, W. Huang, G. Wan, H. Chang, T. Li, Y. Yu, C. Li, J. Li, L. Bai, J. Zhang, Q. Guo, J. Wang, T. Chen, J. T. Zhou, X. Jia, W. Sun, C. Wu, J. Chen, X. Hu, Y. Li, X. Wang, N. Zhang, L. A. Tuan, G. Xu, J. Zhang, T. Zhang, X. Ma, J. Gu, L. Pang, X. Wang, B. An, J. Sun, M. Bansal, S. Pan, L. Lyu, Y. Elovici, B. Kailkhura, Y. Yang, H. Li, W. Xu, Y. Sun, W. Wang, Q. Li, K. Tang, Y.-G. Jiang, F. Juefei-Xu, H. Xiong, X. Wang, D. Tao, P. S. Yu, Q. Wen, Y. Liu, A comprehensive survey in LLM(-agent) full stack safety: Data, training and deployment, 2025. URL: https://arxiv.org/abs/2504.15585. arXiv:2504.15585.

[53] Y. Wang, Z. Fu, J. Cai, P. Tang, H. Lyu, Y. Fang, Z. Zheng, J. Zhou, G. Zeng, C. Xiao, et al., Ultra-fineweb: Efficient data filtering and verification for high-quality LLM training data, arXiv preprint arXiv:2505.05427 (2025).

[54] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al., Training a helpful and harmless assistant with reinforcement learning from human feedback, arXiv preprint arXiv:2204.05862 (2022).

[55] Z. Lin, S. Guan, W. Zhang, H. Zhang, Y. Li, H. Zhang, Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models, Artificial Intelligence Review 57 (2024) 243.

[56] D. H. Lindquist, Y. Dautaj, AI in international arbitration: Need for the human touch, J. Disp. Resol. (2021) 39.

[57] Y. Zhang, M. A. Tuk, A.-K. Klesse, Giving AI a human touch: Highlighting human input increases the perceived helpfulness of advice from AI coaches, Journal of the Association for Consumer Research 9 (2024) 344–356.

[58] B. Porter, E. Machery, AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably, Scientific Reports 14 (2024) 26133.

[59] M. Stanko-Kaczmarek, L. Dera, H. Koscielska, Between the lines: Perceptions of poetry with authorship attributed to artificial intelligence or humans–a comparative analysis, The Journal of Creative Behavior 59 (2025) e1513.

[60] A. Marklová, O. Vinš, M. Vokáčová, J. Milička, The author is dead, but what if they never lived? a reception experiment on czech AI- and human-authored poetry, 2025. URL: https://arxiv.org/abs/2511.21629. arXiv:2511.21629.

[61] J. McCormack, T. Gifford, P. Hutchings, Autonomy, authenticity, authorship and intention in computer generated art, in: International conference on computational intelligence in music, sound, art and design (part of EvoStar), Springer, 2019, pp. 35–50.

[62] L. Switzky, ELIZA effects: Pygmalion and the early development of artificial intelligence, Shaw 40 (2020) 50–68.

[63] J. Milička, Theoretical and methodological framework for studying texts produced by large language models, arXiv preprint arXiv:2408.16740 (2024).