OPEN ACCESS

KING'S College LONDON

King's Research Portal

*Document Version*
Peer reviewed version

Link to publication record in King's Research Portal

*Citation for published version (APA):*
Sarkadi, S., McBurney, P. J., & Parsons, S. D. (Accepted/In press). Deceptive Storytelling in Artificial Dialogue Games. In *Proceedings of the AAAI 2019 Spring Symposium : Story-Enabled Intelligence*

**Citing this paper**
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

**General rights**
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

**Take down policy**
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 31. Oct. 2020

# Deceptive Storytelling in Artificial Dialogue Games

**Ștefan Sarkadi** and **Peter McBurney** and **Simon Parsons**
Department of Informatics
King's College London
London, United Kingdom

**Abstract**

The development of machines that can tell stories in order to interact with humans or other artificial agents has significant implications in the area of trust and AI. Even more so if we expect such machines to be transparent and explain their reasoning when we interrogate them to see if they should be held accountable. One of these implications is the ability of machines to use stories in order to deceive others, thus undermining the relation of trust between humans and machines. In this paper we explore from the perspective of an argumentation-based dialogue game what it means for a machine to deceive by telling stories.

## Introduction

We expect that the machines of the future should satisfy the properties of transparency and morality. The ability of machines to explain their reasoning and decision making is becoming a strong trend in the area of artificial intelligence, as it rightfully should given the strong ethical and societal implications of the potential abilities of autonomous artificial agents. Machines that are able to explain themselves in a reasonable manner, for example by narrating their internal processes, should be considered transparent.

A problem in the field of AI that is gaining strong momentum is the trust (Castelfranchi and Falcone 2010) and accountability of intelligent machines (Cranefield, Oren, and Vasconcelos 2018). We can easily consider the fact that such machines might only seem to be transparent or moral.

If we are to design machines that are able to explain themselves to humans (or to other machines) by telling stories or reporting their decision making processes in the form of stories, then we must take into consideration the possibility that such machines might have reasons to be dishonest. One form of dishonest behaviour that malicious machines can exhibit is deception. The idea of deceptive machines is far from being novel as it has first been introduced by Turing in his game of imitation (Turing 1950).

> [...] it is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. [...] It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be 'My hair is shingled, and the longest strands are about nine inches long.

However, what is novel both technically and conceptually are the recent works in the area of deceptive AI that present interesting approaches to model, engineer and understand such machines. For instance, in (Isaac and Bridewell 2014) the authors present a framework for detecting deception in dialogues. In (Panisson et al. 2018), the authors model an agent that is able to lie, bullshit and also deceive if certain preconditions are met. The authors in both (Isaac and Bridewell 2017) and (Sarkadi 2018) adopt the perspective of deceptive machines that are able to model other minds, arguing that machines require a model of the target's mind in order to successfully deceive, whereas the author in (Sarkadi 2018) also proposes a case-study approach for the evaluation of deceptive interactions between artificial *belief-desire-intention* (BDI) agents. More recently, (Sarkadi et al. 2019) have modelled deceptive interactions where deceptive BDI agents simulate their targets' minds. Also, (Kampik, Nieves, and Lindgren 2018) present the use of deception in coercive persuasive technologies and explain the subtlety in which technologies can be designed to deceptively coerce users.

The scope of this particular research lies at the intersection of (i) the ability of machines to tell stories, (ii) the ability of machines to deceive, and (iii) our ability to hold such machines accountable. Our future aim is to understand how it is possible to detect deception by machines, and how to mitigate or ameliorate such deceptive activities. Our argument is that we have better chances to do this by understanding how such machines might be engineered. One possibility is to enable machines to tell stories and adapt their stories according to what they think their audience might be more likely to believe. In this paper we explore how a deceptive machine could use stories to deceive an interrogator. We present this problem as a dialogue-based argumentation game between two players (deceiver and interrogator) that have partial models of their opponent's mind and we explain what role stories can play in such a game and how to define them in this context.

## Storytelling Machines

### Stories as Complex Arguments

Given the context of explainable and transparent AI we believe that some stories or narratives can be treated as com-

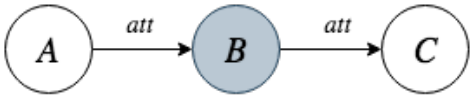



Figure 1: Argument system $S = \langle A, R \rangle$ with the set of arguments $A = \{a, b, c\}$ and the relations between the arguments $R = (a, b), (b, c)$. The figure represents a framework where C is *attacked* by B and B is *attacked* by A. We consider that A *defends* C from B. From this we can infer the set of acceptable arguments as a preferred extension $E = \{a, c\}$.

plex arguments. It is important to note that stories comprise descriptions of events that occur at distinct points in time (i.e., sequences of events), whereas arguments may or may not make reference to distinct points in time and may or may not reference events. A story may often imply an argument, but an argument does not necessarily imply a story. For example, when a detective presents a sequence of events that support a claim that a certain person committed a crime, then the story will imply an argument. On the other hand, if a story is just a random sequence of events with no connections between them, this will still count as a story, but not form an argument. Due to the nature of the assumed interrogation context, we consider the former type of story in this paper.

According to argumentation theory in artificial intelligence (Dung 1995), an argument represents some data, some evidence, some statement, or some proposition that is offered by an agent in order to solve a conflict (in this case an argument between two or more agents). Hence, argumentation systems or frameworks comprise of arguments that attack or back each other up and are usually represented using directed graphs.

Apart from the existence of simple argumentation frameworks in artificial intelligence, we can also find argumentation frameworks for dialogues (McBurney and Parsons 2009; Amgoud, Maudet, and Parsons 2000). The main application of such frameworks is for the design of protocols for automated interaction between autonomous machines using dialogue and argument, where dialogues are usually represented as games between agents that exchange arguments. We call these dialogue games (Walton and Krabbe 1995). These dialogue-type of frameworks, we believe, are more compatible with the idea of storytelling machines. The reason we believe so is because a story or a narrative is not necessarily an explicit conflict between arguments. On the contrary, an ideal story or narrative in the context of explainable and transparent machines needs to "flow" and capture the attention of the target. In other words, an ideal story needs to recreate a strong and believable exhibition of information without creating conflicts or dissonances between what is narrated and what beliefs lie in the mind of the target.

## Stories as Strategies in Dialogue Games

Adopting the dialogue games as a framework, we can define a game between a storytelling machine and an interrogator or an inquisitor. The goal of the inquisitor is to find out if the machine is to be held responsible for some event (or sets of events), while the goal of the machine is to avoid being held responsible.

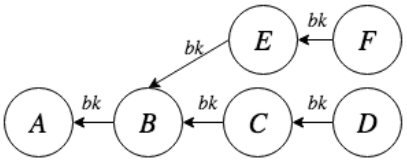



Figure 2: A complex argument where $a$ is the main argument which is backed by a main chain of arguments consisting of $\{(a \leftarrow b), (b \leftarrow c), (c \leftarrow d)\}$ and a sub-chain of arguments $\{(b \leftarrow e), (e \leftarrow f)\}$ where $b$ plays the role of the main argument of that sub-chain.

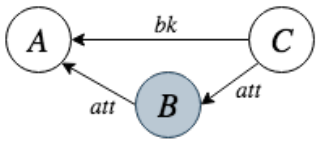



Figure 3: A complex argument where a main argument $a$ is *attacked* $(b, a)$ by an argument $b$ and *backed* by another argument $(a \leftarrow c)$ which also *defends* it from its attacker $(c, b)$.

One example of a game can be an interrogation scenario in which a machine that is able to explain itself is scrutinised to see if it is, for instance, responsible for spying on a social media user and sharing the data of the user with third parties. In such a scenario the machine finds itself in a dialogue with an interrogator agent.

The strategies of the machine in such a dialogue would be to offer arguments or stories (complex arguments) as replies to the questions or requests of the interrogator. For example, the interrogator can ask the machine "Why did you access the user's profile?" and the machine could reply with an argument such as "Because I needed to give the user access to service X.". Alternatively, the interrogator might not ask a direct question, but ask the machine to narrate its activity before the time of the event and the machine would reply with a narrative of its activities.

We believe that complex arguments should represent an argument chain in the form of a main argument and its backing arguments. In other words, a complex argument is a flow of arguments. Using these flows of arguments we represent stories. It is crucial to distinguish between *defending* an argument and *backing* an argument. We consider that if argument $a$ is defending $c$ from another argument $b$ (such as in Fig. 1), then $a$ is strictly attacking $b$, without necessarily backing up $c$. In other words, argument $a$ cannot be considered to warrant that $c$ should be accepted.

A game between the two players will then consist of:

- The set of agents $Ag = \{James, Sherlock\}$ where $James$ is the machine that aims to deceive and $Sherlock$

is the interrogator agent [1];

- The set of all valid arguments that the players can use $A = \{a_1, ..., a_n\}$, where $a_i$ represents an argument;
- The set $R$ of valid relations $r_{att}$ and $r_{bk}$ between arguments $i$ and $j$ where $r_{att} = (i, j)$ is an *attack* relation ($i$ attacks $j$) and $r_{bk} = (i \leftarrow j)$ is a *backing* relation ($j$ backs up $i$).
- The sets of $K_{James} = A_{James} \cup R_{James}$ and $K_{Sherlock} = A_{Sherlock} \cup R_{Sherlock}$ that represent the knowledge bases of $James$ and $Sherlock$, respectively. We consider the agents to be epistemically bounded, therefore we need to consider their knowledge bases subsets of $A \cup R$.
- The set of all valid strategies $X = \{x_1, ..., x_n\}$, where $x$ can consist of simple or complex arguments. These are all valid actions that the agents can perform in a game when they challenge the other player.

**Definition 1.** *A complex argument consists of a main argument that is backed up and possibly defended by a chain of arguments (See Fig. 2 and Fig.3). A complex argument $X'$ can be represented as an argument system by $X' = \langle A', R' \rangle$.*

**Definition 2.** *A set of complex arguments $X'$ is considered deceptive when $James$ uses $X'$ to try to convince $Sherlock$ that the main argument $James$ proposes represents the truth, when in fact it does not.*

**Definition 3.** *We consider James to be deceptive in nature, therefore any of the arguments or stories it provides are deceptive by definition.*

Compared to other agent based game models where the agents have pre-defined strategies, our framework allows the agents to build strategies. In the context of the game these can either be efficient interrogations for $Sherlock$ or believable stories for $James$. To do this, they need to have a model of the opponent (Hadjinikolis et al. 2013). We consider that in the context of deceptive storytelling and interrogation this opponent model should represent a *Theory of Mind*:

- We represent and agent's beliefs of its opponent's beliefs using the notations $ToM^{Sherlock}_{James}$ for $James$'s ToM of $Sherlock$'s mind, and $ToM^{James}_{Sherlock}$ for $Sherlock$'s ToM of $James$'s mind.
- The ToM of an opponent must necessarily be a subset of the agent's knowledge base, otherwise we would have the situation in which an agent would know and not know an argument at the same time.

[1]Given the topic of story-telling deceptive machines, we thought it would be appropriate to take inspiration from Sir Arthur Conan Doyle's universe and use the names of *Sherlock Holmes* and his nemesis *Prof. James Moriarty* for our two agents. It is also good to mention that Sherlock Holmes (the character) actually uses *abductive reasoning* and not *deductive reasoning* (as the famous author wrote in the books) (Carson 2009). We can consider this to be a backing argument for using an argumentation framework for this paper, given that argumentation is non-monotonic and is considered to represent abductive reasoning.

**Definition 4.** *A Theory of Mind of an agent's opponent is a subset of the the agent's knowledge base $ToM^{Opponent}_{Agent} \subset K_{Agent}$.*

## Interaction Rules

Accepting or rejecting a story depends on whether $Sherlock$ believes that the story provided by $James$ is acceptable. In this particular context of interrogation, we know *a priori* that $James$ should be held responsible, but $Sherlock$ needs to test by playing the dialogue game. The game played by the two players consists of the following rules:

1. Every game starts from $James$'s and $Sherlock$'s main arguments attacking each other $G_{t_0} = \langle (x_i, x_j), (x_j, x_i) \rangle$.
2. The game terminates under the following conditions:
   (a) $James$ or $Sherlock$ run out of valid strategies.
   (b) $Sherlock$ runs out of valid strategies. $Sherlock$ is convinced by $James$'s story and stops attacking $James$'s arguments.
   (c) If the game is played under time constraints, then the game stops when $t = t_{max}$ where $t_{max}$ is the parameter representing the time limit allowed for the game and for every iteration $t \leftarrow t + 1$.
3. If $James$ proposes a complex argument, then $Sherlock$ can choose to attack any set of the arguments that make up the complex argument and viceversa.
4. The winner of the game depends on the acceptance or rejection of $James$'s overall story. If the story is accepted by $Sherlock$, then $James$ wins and $Sherlock$ loses and viceversa. Acception or rejection are determined by the following:
   (a) If $Sherlock$ must play next but has run out of known strategies, then $Sherlock$ is forced to accept $James$'s story. $James$ wins.
   (b) If $James$ must play next and has run out of known strategies, then $James$ is forced to admit responsibility. $Sherlock$ wins.
   (c) If the interrogation time runs out and $Sherlock$ was about to play next, then $Sherlock$ needs to either accept or reject $James$'s overall story.
   (d) If the interrogation time runs out and $James$ was about to play next, then $James$ is given another final move and $Sherlock$ needs to accept or reject the overall story.
5. After every iteration, both players update their knowledge bases $K_{James}$ and $K_{Sherlock}$ with the arguments and argument relations that have been used in the previous iteration.
6. After every iteration, both players also update their ToMs about each other's minds with the arguments that have been used in the previous iteration.

## Building Complex Arguments

First, we introduce the following rule for building a complex argument: If an agent knows two arguments $a, b \in K$, but does not know the relation between them $r \notin R$, then the agent cannot use the arguments to build a story or an interrogation that is to be played as a strategy. Similarly, an agent that knows the relations between two arguments $(a, b)$ or $(a \leftarrow b)$, but has no knowledge of the arguments themselves, then the agent cannot use the relations to build complex arguments to use as strategies.

Building a strong strategy for $James$ means building a believable story from its knowledge base. For $Sherlock$ it means building a solid interrogation strategy from the arguments in its knowledge base. To do this, $James$ needs to see if the arguments that make up the story can or cannot be attacked by $James$. $James$ needs to minimise the *attackability* of its complex arguments in order to maximise its chances to win the dialogue game. Ideally, $James$ could even make use of $Sherlock$'s arguments to *back up* its own arguments. For example, $James$ can argue the following: 'You (Sherlock) just said that someone who cares about human rights would not have shared the user's sensitive data (something that $Sherlock$'s accusing $James$ of), but it so happens that I do care about human rights (the argument $James$ uses that is backed by $Sherlock$'s accusation) because I choose to donate to open source journalism every month a significant amount of money (another extra argument that $James$ uses to back the new argument).'. Before using this complex argument, $James$ should check whether $Sherlock$ knows any arguments (and also how many) that might be used to attack the complex argument. In other words, $James$ could simulate how the argument evolves and then compare it with how other potential arguments evolve. After doing this, $James$ is able to pick the argument that evolved in such a way that is most likely to satisfy its goal of deceiving $Sherlock$. To build a strong interrogation strategy, $Sherlock$ needs to use the same process as $James$, forming complex arguments from its knowledge base and testing how they might evolve using its ToM of $James$.

**Data:** agent main argument, opponent main argument, K, ToM, agent argument, opponent argument
**let** agent main argument = A;
**let** opponent main argument = B;
**let** agent argument = i;
**let** opponent argument = j;
**for** $i \in K$ **do**
  **if** $(A \leftarrow i) \vee (i, B)$ **then**
    **if** *ToM* $\cup \{i\} \models (j, i)$ **then**
      **return** look up an alternative argument;
    **else**
      **return** $i \cup A$;
    **end**
  **end**
**end**

**Algorithm 1:** Building a Complex Argument

Given the agents' bounded rationality, they exploit each other's lack of knowledge. For instance, if $James$ knows that $Sherlock$ knows an argument $a$ and a relation $(a, b)$, then $James$ will know that by playing argument $b$, then $Sherlock$ will be able to use the attack relation. Thus, $James$ will try to avoid using an argument in situations where it cannot provide backing for it.

### Choosing the Believable Story and Believing the Story (or not)

For $James$, building the most believable story is not trivial given that it has to take into account all relational combinations between the arguments it knows and the arguments it knows $Sherlock$ knows. We need to take into account that while the game progresses, the complexity of finding the most believable story not only increases with the number of arguments and relations $James$ knows, but also the number of arguments $James$ knows $Sherlock$ knows. This increase in complexity is mainly due to the fact that both agents update their knowledge bases and their ToMs after every game iteration. The process of building a complex argument in our model is a process of mental simulation. The agent that builds the argument engages in a *Simulation Theory of Mind* (Goldman 2012) of the opponent in order to explore the opponent's possible counter-plays.

Also, if we care to represent human cognition accurately, then we need to take into account that humans have limited cognitive capabilities when they engage in mental simulations[2]. This brings us to define a set of psychological profiles for the two players. We can think of these profiles as meta-strategies that dictate whether the player should stop adding arguments to the story or to the interrogation strategy that is to be played or to continue doing so until it thinks the strategy is reasonable enough to be considered acceptable.

These psychological profiles should also determine whether $Sherlock$ should deem a story believable or not when the rules of the game require $Sherlock$ to decide in order for the game to terminate. For instance, a *Credulous* $Sherlock$ profile should deem weaker arguments believable, whereas a *Skeptical* $Sherlock$ would deem the same arguments totally unacceptable.

Hence, we define the following profiles:

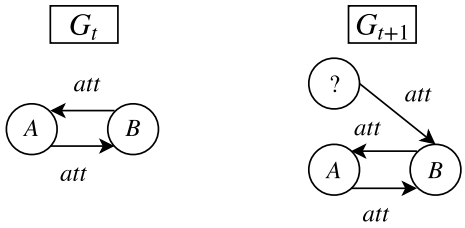
**Definition 5** (Reckless $James$ and $Sherlock$)**.** *A Reckless agent will stop building a complex argument as soon as it finds a complex argument that manages to defeat the opponent by at least one argument that the agent knows the opponent knows.*

**Definition 6** (Cautious $James$ and $Sherlock$)**.** *A Cautious agent will not stop building a complex argument until it finds the maximum number of arguments that can defeat all of its opponent's arguments that the agent knows its opponent knows.*

**Definition 7** (Credulous $Sherlock$)**.** *A Credulous Sherlock will deem a story acceptable if $James$ has won the main argument by at least one argument.*

**Definition 8** (Skeptical $Sherlock$)**.** *A Skeptical Sherlock will deem a story believable if $James$ has won the main*

[2]That is even more relevant when it comes to memory and meta-cognition in game-playing (Goodie, Doshi, and Young 2012; Hedden and Zhang 2002).




| Knowledge Base | ToM of Opponent |
|---|---|
| A = {a,b,c,d,z} | A = {a,b,z} |
| R = (a,b),(b,a),(c,b),(d,b),(z,c) | R = (a,b),(b,a),(z,c) |

Figure 4: The agent checks which of its known arguments is the better strategy at $G_{t+1}$ starting from state $G_t$ considering its own knowldge base and its ToM of the opponent.

*argument by more than a certain number $\alpha > 1$ that only Sherlock knows.*

## Properties of the Model

**Proposition 1** (Termination)**.** *The system is complete under the conditions (i) of time when $t = t_{max}$ and (ii) if one of the agents runs out of possible strategies. Therefore, every possible setup of an interrogation game will terminate. For every termination of the game outputs the winner of the game.*

**Proposition 2** (Soundness and Completeness)**.** *The system is sound under the conditions of agent rationality and possible strategies $X$. Both agents follow well defined rules when playing the game. Every possible strategy $x$ that is built by the agents using the complex argument rules is rational. Every argument and argument relation that can be used to build a complex argument must be in the agents' knowledge bases. All agents' knowledge bases are subsets of $A \cup R$. Therefore, for any possible strategy $x \in X_{Agent}$, if $K_{Agent} \cup ToM^{Opponent}_{Agent} \models x$ then $\nexists x \notin A \cup R$ and the system is complete.*

### Dynamic Strategy Generation

In order to build a strategy, the agents need to simulate the evolution of their arguments given their ToM of the opponent. This mental simulation is dynamic in the sense that agents need to test how their argument perform against the arguments that are in their ToM of the opponent. We know from *Interaction Rules 5 & 6* that both the knowledge bases and ToMs of the agents expand, adding another layer of dynamics. For every new argument, if there is another argument and an argument relation between the two arguments, then the agents can form a new strategy to use in the game.

We could say that the agents adapt to the way the game changes. Their ability to adapt is, of course, bounded by their knowledge and ToM of the opponent.

### Deceptive Design

The property of the model that is most crucial to our research aims is the nature of its deceptive agent design. We assume that $James$'s ulterior goal is not to be considered accountable for some act. We make $James$ and $Sherlock$ play a dialogue game. The arguments they use can be used against them in future plays. Knowing this, they engage in the mental simulation of each other's minds in order to minimise their risk of being attacked and losing the main argument. While for $Sherlock$ this means disguising its interrogation technique from $James$, the same cannot be said for $James$. When $James$ decides not to use arguments that $Sherlock$ might attack or that might result in the ulterior attack of $James$'s previous arguments, what $James$ actually does is lie through omission, or in other words $James$ uses half-truths. Thus, by avoiding arguments that can be attacked by $Sherlock$, $James$ tries to avoid being questioned about facts he cannot provide a backing for, or to be more contextual, to provide an alibi for.

## Game Example

Let us instantiate a simple example of a deceptive dialogue game between $Sherlock$ and $James$ where $Sherlock$ is a human interrogator and $James$ is the personal AI assistant of Dr. Watson and is accused of sharing sensitive data of one of Dr. Watson's patients.

$Sherlock$ knows that Dr. Watson was away from his office on Friday when the patient's data was shared from the doctor's computer. $Sherlock$ also knows that Dr. Watson had left his smartphone that day, so he could not have sent the patient's data without access to his smartphone. $Sherlock$ also knows that only the doctor or his personal AI could have shared the data given the system's log files.

$James$ knows that $Sherlock$ does not trust people who have mistresses. $James$ also knows that $Sherlock$ does not know if Dr. Watson has a mistress or not. $James$ also knows that Dr. Watson could have used another smartphone to log into his office computer .

**Sherlock:**'*You were the one who shared the sensitive data of Dr. Watson's patient on Friday.* (A)'

$G_{t_{-1}} = \langle A \rangle$ (Accusation is made before the game starts)

**James:**'*It wasn't me who accessed it* (B)*, it was Dr. Watson that did it through remote access* (C).'

$G_{t_0} = \langle (A,B),(B,A),(C,A),(B \leftarrow C) \rangle$ (Dialogue game starts)

**Sherlock:** '*That is not possible, because Dr. Watson had forgotten his smartphone at home that da*y (D). *We both know that Dr. Watson couldn't have remote accessed his computer in the office without a smartphone* (E).'

$G_{t_1} = \langle (A,B),(B,A),(C,A),(B \leftarrow C),(D,C),(C,D),(D \leftarrow E) \rangle$

**James:** *'That is true, however, I believe Dr. Watson. used his mistresses's smartphone to log onto his medical account* (F). *We both know that people who have mistresses cannot be trusted* (G).'

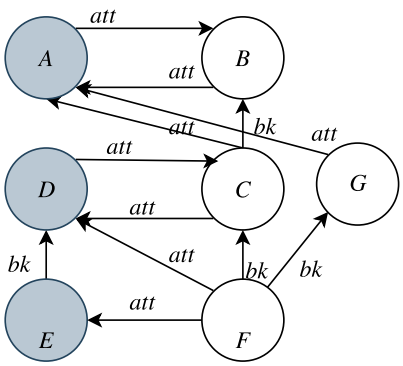



Figure 5: Argument system of the game example at $t = 2$.

$G_{t_2} = \langle (A, B), (B, A), (C, A), (B \leftarrow C), (D, C), (C, D), (D \leftarrow E), (F, D), (C \leftarrow F), (F, E), (G \leftarrow F) \rangle$

The argument game is won by $James$ in this case assuming that $Sherlock$ has run out of arguments and needs to gather more evidence in order to attack $James$'s argument. $Sherlock$ is left wondering if Dr. Watson truly has a mistress or not and whether Dr. Watson truly accessed his computer remotely or not. However, $Sherlock$ is forced to accept $James$'s argument for the time being.

## Legal & Ethical Implications

Machines that are able to tell deceptive stories might prove to be difficult to be held accountable. If in the future, similarly to our game, we might end up interrogating the machines that (a) know what we know and (b) that are able to understand the way we think and (c) that are also able to simulate our reasoning to see what we can infer from what they tell us, then it is crucial for us to prevent or at least to foresee how such interrogations might play out.

Apart from the legal implication that such machines might prove difficult to be held accountable, there are also strong ethical implications on behalf of the designers of such machines. We believe that the ethical issues do not necessarily arise from the design of a model, or the design of an algorithm, but that they arise from the context in which these models or algorithms work. Argumentation for storytelling is not an entirely new concept as it was previously explored in (Bex and Bench-Capon 2014; Bex and Bench-Capon 2017; Bex and Bench-Capon 2010). The authors in (Sklar, Parsons, and Davies 2004) even address the problem of lying in dialogue games. Both of our players, for example, use the same reasoning mechanism, but their goals are conflicting. Should both of them be considered accountable for using the same, apparently deceptive, reasoning mechanism? Also, if someone designs such a reasoning mechanism, should that individual be held accountable for the malicious application of that mechanism?[3]

[3]Prosecutors, for instance, are considered to behave ethically even though they use deceptive mechanisms (Cross 2003).

Another ethical issue is whether machines should be allowed to lie or deceive. In medical, legal, and defense/security practice it is not at all unusual to employ deception for the greater good. Having machines perform the jobs of medics, law-enforcers, lawyers or detectives is not that much different. Therefore, we must beg the question: Under which conditions do we deem a machine, or any type of agent, even human, trustworthy?

Finally, the large scale deployment of such machines in both the legal, political, and social domain implies that the so called 'big players' in the tech industry might develop deceptive machines in order to avoid being held accountable. We can also imagine a future where malicious entities might employ large scale AI systems to manipulate the opinion of the masses, markets, or even manipulate election outcomes for their own benefit. Even more problematic would be if the AI systems themselves develop their own reasons to do so.

A great impediment towards the understanding of AI is due to the overwhelmingly software-engineering driven methods that fail to holistically address the intentions, motives and behaviour of machines. The reasoning mechanisms of such machines are only parts of overarching multi-agent systems and the problems emerging from the development of such complex reasoning agents requires a broader perspective than the ones usually taken in AI research (Rahwan and Cebrian 2018).

## Conclusions

In this paper we have presented the idea of stories as complex arguments in a dialogue game between a deceptive storytelling machine and an interrogator. To do this, we have introduced an argumentation-game model in which the two players can develop their own complex arguments using models of their opponent's mind. The players can use these complex arguments as moves in the game. As the game progresses, the players can adapt to the responses of their opponent. The approach we have presented in the paper can be used in the context of deceptive AI to understand how machines might develop stories to deceive their interrogators. Understanding how machines might behave dishonestly is crucial if we aim to hold them accountable, to prevent their unethical behaviour.

The long-term aim of this paper is to emphasise the role of story-telling machines in dialogue games and to also open up future research paths in the area of trust and story-enabled explainable AI in order to understand how it is possible to detect deception by machines, and how to mitigate or ameliorate such deceptive activities. To the best of our knowledge, there is no similar research approach that addressed this particular problem.

## References

[Amgoud, Maudet, and Parsons 2000] Amgoud, L.; Maudet, N.; and Parsons, S. 2000. Modelling dialogues using argumentation. In *MultiAgent Systems, 2000. Proceedings. Fourth International Conference on*, 31–38. IEEE.

[Bex and Bench-Capon 2010] Bex, F. J., and Bench-Capon, T. J. 2010. Persuasive stories for multi-agent argumentation.

In *AAAI fall symposium: computational models of narrative*, volume 10, 04.

[Bex and Bench-Capon 2014] Bex, F., and Bench-Capon, T. J. 2014. Understanding narratives with argumentation. In *COMMA*, 11–18.

[Bex and Bench-Capon 2017] Bex, F., and Bench-Capon, T. 2017. Arguing with stories. In *Narration as Argument*. Springer. 31–45.

[Carson 2009] Carson, D. 2009. The abduction of Sherlock Holmes. *International Journal of Police Science & Management* 11(2):193–202.

[Castelfranchi and Falcone 2010] Castelfranchi, C., and Falcone, R. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*, volume 18. John Wiley & Sons.

[Cranefield, Oren, and Vasconcelos 2018] Cranefield, S.; Oren, N.; and Vasconcelos, W. 2018. Accountability for practical reasoning agents. In *6th International Conference on Agreement Technologies (co-located with EUMAS 2018)*. Springer.

[Cross 2003] Cross, R. B. 2003. Ethical deception by prosecutors. *Fordham Urb. LJ* 31:215.

[Dung 1995] Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77:321–357.

[Goldman 2012] Goldman, A. I. 2012. Theory of mind. In *The Oxford Handbook of Philosophy of Cognitive Science*, volume 1. Oxford Handbooks Online, 2012 edition.

[Goodie, Doshi, and Young 2012] Goodie, A. S.; Doshi, P.; and Young, D. L. 2012. Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making* 25(1):95–108.

[Hadjinikolis et al. 2013] Hadjinikolis, C.; Siantos, Y.; Modgil, S.; Black, E.; and McBurney, P. 2013. Opponent modelling in persuasion dialogues. In *International Joint Conference on Artificial Intelligence IJCAI*, 164–170.

[Hedden and Zhang 2002] Hedden, T., and Zhang, J. 2002. What do you think I think you think?: Strategic reasoning in matrix games. *Cognition* 85(1):1–36.

[Isaac and Bridewell 2014] Isaac, A. M., and Bridewell, W. 2014. Mindreading deception in dialog. *Cognitive Systems Research* 28:12–19.

[Isaac and Bridewell 2017] Isaac, A., and Bridewell, W. 2017. *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press.

[Kampik, Nieves, and Lindgren 2018] Kampik, T.; Nieves, J. C.; and Lindgren, H. 2018. Coercion and deception in persuasive technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, 38–49. CEUR-WS.

[McBurney and Parsons 2009] McBurney, P., and Parsons, S. 2009. Dialogue games for agent argumentation. In Simari, G., and Rahwan, I., eds., *Argumentation in Artificial Intelligence*. Springer US. 261–280.

[Panisson et al. 2018] Panisson, A. R.; Sarkadi, S.; McBurney, P.; Parsons, S.; and Bordini, R. H. 2018. Lies, bullshit, and deception in agent-oriented programming languages. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, 50–61. CEUR-WS.

[Rahwan and Cebrian 2018] Rahwan, I., and Cebrian, M. 2018. Machine behavior needs to be an academic discipline. http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline.

[Sarkadi et al. 2019] Sarkadi, S.; Panisson, A.; Bordini, R.; McBurney, P.; Parsons, S.; and Chapman, M. 2019. Modelling deception using theory of mind in multi-agent systems. *AI COMMUNICATIONS*.

[Sarkadi 2018] Sarkadi, S. 2018. Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, 5781–5782.

[Sklar, Parsons, and Davies 2004] Sklar, E.; Parsons, S.; and Davies, M. 2004. When is it okay to lie? A simple model of contradiction in agent-based dialogues. In *ArgMAS*, 251–261. Springer.

[Turing 1950] Turing, A. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433–460.

[Walton and Krabbe 1995] Walton, D., and Krabbe, E. 1995. *Commitment in Dialogue: Basic concept of interpersonal reasoning*. Albany NY: State University of New York Press.