ExplOrs

Explanation Oracles and the architecture of explainability

Ravikanth Pappu

May 19, 2020

Abstract

We introduce Explanation Oracles (hereafter *ExplOrs*) in this paper. In response to an input and one or more questions about that input, they generate and convey *conversational* explanations to humans. ExplOrs have access to domain-specific external information that enables them to reason about the input. They operate in an adversarial setting - some of the inputs they receive are designed to deceive them. The ExplOr attempts to detect this deception with probability close to unity and explain it. The recipient of the explanation has a binary choice of accepting or rejecting the explanation. This *human binarization* reduces the problem of evaluating explanations into a decision problem, and sidesteps the sensitive dependence of explanations on the input, queries, and models used in the system. ExplOrs maintain a record of unique inputs, corresponding outputs, and the human's decision for use in the future. Thinking about explainability in this setting yields a prescription for architecting Explainable AI systems (XAI) in complex domains. In this paper, we ask what capabilities an ExplOr needs to successfully explain its conclusions to human experts. We are solely interested in the *architecture* of ExplOrs, not in specific implementations.

1 Introduction

Oracles are, according to the New Oxford American dictionary, *"a person or thing regarded as an infallible authority or guide on something"*. The concept of an oracle is widely used in cryptography. For instance, the Random Oracle Model (ROM) [6] is an idealized version of a perfect cryptographic hash function. When presented with an input string S, the oracle instantaneously returns a fixed width random string $f(S)$. Oracles are a useful abstraction because they enable the design of and reasoning about cryptographic protocols without getting bogged down in the details of specific implementations.

Explainable AI (XAI) [49, 12, 28, 8] is a varied and complex problem and there is no agreement on what constitutes explainability or what that term even means. Among scores of others, questions ranging from: *how did this Deep Neural Network decide there is a cat in this image?* to *why was a loan denied* *to a specific applicant?* are included in the realm of explainability. From the literature, it is clear that the field is far from agreement on a common vocabulary, definitions, and what constitutes an explanation. However, as AI systems become more pervasive, especially in regulated environments like health and safety, there is a growing need to formalize explainability and the performance of XAI systems.

We introduce Explanation Oracles in this paper and argue that thinking about explainability in this setting yields a prescription for architecting Explainable AI systems (XAI) in complex domains. In response to an input (which is a multi-dimensional array in the most general case) and one or more queries about that input, ExplOrs generate and convey *conversational* explanations [15, 18] to humans. ExplOrs have access to domain-specific and external information that enables them to reason about the input. They operate in an adversarial setting - some of the inputs they receive are designed to deceive them. The ExplOr attempts to detect this deception with probability close to unity and explain it. The recipient of the explanation has a binary choice of accepting or rejecting the explanation. This *human binarization* reduces the problem of evaluating explanations into a decision problem, and sidesteps the sensitive dependence of explanations on the input, queries, and models uses in the system. ExplOrs maintain a record of unique inputs, corresponding outputs, and the human's decision for use in the future. In this paper, we ask what capabilities an ExplOr needs to successfully explain its conclusions to human experts. We are solely interested in the *architecture* of ExplOrs, not in specific implementations.

By clearly defining a system boundary for ExplOrs, making the role of the human explicit, and encapsulating their functionality, we afford ourselves the opportunity to compare different explanation approaches against each other, discover open questions, and chart improvement in XAI in specific domains over time.

This paper is divided into two parts. *The magic of explanation* describes how explanations are structured in the domain of magic tricks. Although this part is long and relegated to the appendix, it is an integral part of the paper, and the most entertaining, so the reader is urged to read it and watch the referenced videos. If you are short on time, *read section A.3* at a minimum. *The architecture of explainability*, starting with section 2 below describes our inspiration, related work, formalizes the model of an ExplOr, and introduces the capabilities an ExplOr needs to generate explanations acceptable to a human.

2 Inspiration for ExplOrs

Penn & Teller: Fool Us - an American television show - inspired the idea of ExplOrs. The hosts - Penn Jillette and Teller (hereafter referred to as P&T) - are renowned and highly-experienced magicians who have performed magic together for several decades. On the show, expert magicians from all over the world are invited to perform a magic trick for P&T. If neither host can explain how the trick was done to the satisfaction of the magicians and an expert behind-the-scenes judge, the magicians have successfully fooled P&T. As a prize, they receive a trophy and an all-expenses-paid trip to perform the opening act at P&T's headline show. For clarity, we restrict the meaning of the word *fool* to simply mean that P&T did not successfully explain how the trick was done. Figure 1 depicts a scene from the show.

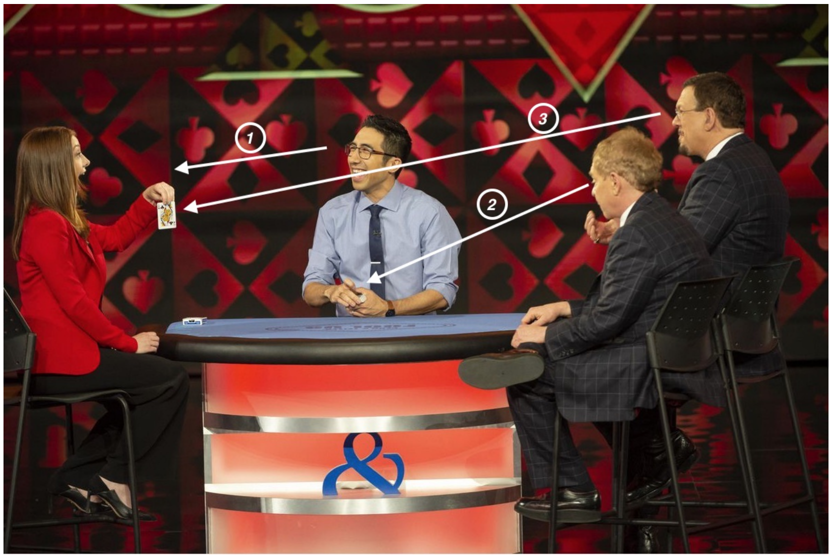

Figure 1: A scene from a performance on *Penn and Teller:Fool Us.* From left to right: Emcee Alyson Hannigan, the performer Jimmy Ichihana doing a close-up card trick, and Teller and Penn observing the performance. (1) Ichihana gets Hannigan to reveal her card eliciting utter delight. (2) Teller is paying unwavering attention to Ichihana's hands. (3) Penn is looking at the result of the effect as well.

This is a show about *generating conversational explanations* in an *adversarial setting* in a *specific domain of expertise*. By dissecting how P&T explain magic tricks, we discern several capabilities that an ExplOr must possess to successfully explain its conclusions to humans. The show discussed in detail in appendix A.

Finally, while we are attempting to use a magic show to inform the design of Explainable AI systems, we note that AI has recently been employed to design magic tricks [47, 48].

3 Related work

This section is not intended to be an exhaustive review of the use of oracles in AI or XAI. Rather, we aim to provide a (very brief) flavor for work that is intellectually similar to our proposal.

At one end of the spectrum of use, oracles are described as a way to sandbox AI in [7] and [3]. To prevent a superintelligent AI from getting out of control, an oracle mechanism is used to limit AIs from acquiring too much information from the external world or affecting it in malicious ways.

Pezeshkpour et al. [31] use a GAN to generate user-friendly explanations of loan denials delivered to loan applicants (i.e., non-experts). Their focus is on the human-understandable aspect of explanations. This is an important aspect of generating conversational explanations.

Rauschecker et al. [34] describe the performance of an AI system designed to generate differential diagnoses of common and rare brain disease by looking at MRI scans of patients. By using a combination of deep learning for brain lesion detection and characterization and naive Bayesian model for differential diagnosis generation, they were able to show that the AI system approaches human expert performance. Domain knowledge was encapsulated in the prior probabilities of the Bayesian model as well as in the lesion detection and characterization. The "explanation", delivered to experts, was simply a ranked list of diagnoses.

Yi et al [50] describe a system that focuses on extracting temporal and causal structure underlying videos of simple objects interacting via Newtonian mechanics. In their system, they use several different techniques for input video understanding, visual question answering, and a new oracle technique - Neuro-Symbolic Dynamic Reasoning - to predict the dynamics of the objects in the videos. Users interact with the system by asking questions. Answers to these questions can be descriptive (e.g., how many metal objects are moving?), explanatory (e.g., which shape is responsible for causing the green sphere to move?), predictive (e.g., what will happen next?), and counterfactual (e.g., which of the following will happen if the gray object is absent?). In this case, the explanation is an answer to a question formulated to narrow down possible answers. The domain and all the options in the system are fixed and carefully controlled.

We observe that in the latter three cases the domain (e.g., finance, neurology, physics) and the attendant vocabulary/ontology of that domain is implicitly contained in the models. We consider [34] and [50] to be instances of ExplOrs but without clearly defining all the implicit assumptions inherent in the systems. We address this challenge next.

4 The architecture of ExplOrs

We now turn our attention to an idealized model of an ExplOr. See Figure 2 for a cartoon of the setting. We start with some definitions and notation.

- Domain D: ExplOrs operate in a specific well-defined domain D. It may be broad or narrow, but intuitively, narrow domains offer a greater chance of successful explanations. The domain has at least one ontology that defines and relates fundamental concepts in the domain to one another. We will assume that the ontology is represented by a labeled directed multigraph which enables reasoning over it.

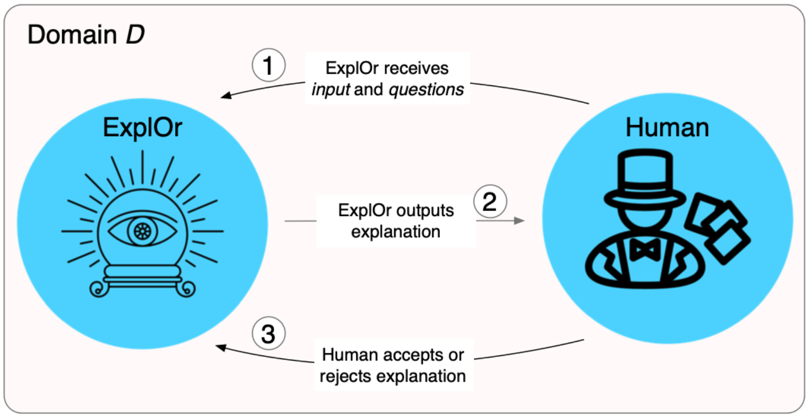
Domain D
1
ExplOr receives *input* and *questions*
ExplOr
2
ExplOr outputs explanation
Human
3
Human accepts or rejects explanation

Figure 2: Cartoon model of the setting considered in this paper. The ExplOr and the human are embedded in a specific domain D. In this context, the ExplOr receives input and queries from a human, answers the questions, and the human either accepts or rejects the explanation.

- Input $T_A^k, k \in \{1, 2, 3, ...N\}, A \in \{0, 1, X\}$: In the most general case the input is a multi-dimensional array. It could be a vector of values of features or a video with an audio track. The input is composed of objects from the domain D. Each distinct input is uniquely identified with an index k so the ExplOr can keep track of unique inputs. The subscript $A \in \{0, 1, X\}$ is an indicator of whether or not the input is not adversarial, adversarial, or unknown. When provided, this lets the ExplOr know whether or not the input is adversarial. This requires an increasing amount of storage in the ExplOr over time.

  Inputs are allowed to be adversarial. Why allow adversarial inputs at all? Adversarial inputs are pervasive. The greater the generalizability demanded of our models, the greater the susceptibility to adversarial input. According to [20], "attaining models that are robust and interpretable will require explicitly encoding *human priors* into the training process." In [14], the authors create medical images which are obviously impossible (e.g., an image of a diseased retina with a scaled image of a normal retina somewhere in the picture) which is classified correctly as being diseased. In [10] and [55] inputs are deliberately authored to target weaknesses of systems. We want an ExplOr to be able to detect and explain these kinds of adversarial or implausible inputs.

- Queries $Q_i^k, k \in \{1, 2, 3, ...N\}, i \in \{1, 2, 3, ...M\}$: This is a set of M questions that the human asks the ExplOr about input T_A^k. Without loss of generality, we can assume that they are strings of text. For the case of the magic show, the single question is *Did I fool you*? In [34] the question is *what is the differential diagnosis consistent with the MRI and patient metadata*? Unique queries are also stored in the ExplOr and accumulate over time.

- Responses $R_i^k, k \in \{1, 2, 3, ...N\}, i \in \{1, 2, 3, ...M\}$: Each query Q_i^k receives a response. Without loss of generality, this is a text string as well. As before, unique responses are stored. We note that acceptable responses can vary in content and that each combination of an input and query could have multiple acceptable responses. We note that we require responses to be conversations which could be interactive.
- Acceptance $A_i^k \in \{0, 1\}, k \in \{1, 2, 3, ...N\}, i \in \{1, 2, 3, ...M\}$: Finally, when the human sees an explanation, they accept or reject it. A_i^k is a Boolean variable that represents this decision - 0 represents an acceptable explanation and 1 represents an unacceptable one. This binarization process allows the existence of multiple acceptable explanations.

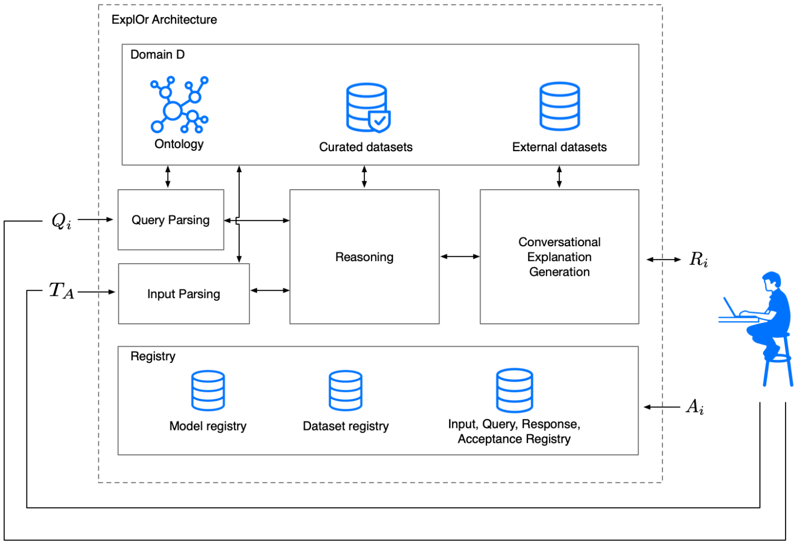

Figure 3: A sketch of the high-level architecture of an ExplOr. The dotted line clearly delineates the system boundary of the ExplOr and the role of the human is made explicit.

Figure 3 shows the high-level architecture of an ExplOr. It lives in a well-defined Domain D with an attendant ontology, a set of curated datasets about that domain, and access to external data about the world.

The *Query Parsing* module receives input queries, parses them, and makes them available to the *Reasoning* module. Similarly the *Input Parsing* module receives the input, detects and flags adversarial and out-of-distribution input. It stores the input/query pair in the registry and passes the parsed input to the *Reasoning* module, which is responsible for determining the answer to the queries given the input. It has access to all the available information in the domain D. Input and query parsing will rely on a variety of deep learning models to achieve their goals. For instance, video input could be parsed with a Mask R-CNN [17] while queries could use an attention-based seq2seq model [4]. To ensure that everything an ExplOr does is traceable, the versions of the models used and the data they were used to train on need to be recorded.

The Reasoning module is responsible for generating the explanation. Explanations are computed in many ways depending on the domain and the queries.. They can range from LIME [35], Shapley values [24], and self-explaining neural networks (SENN) [1] to novel variations on decision trees [42]. [29] provides comprehensive reference on model intepretability, which is one element of an explanation. More advanced ExplOrs will also use other forms of reasoning including causal, analogical, deductive, or inductive reasoning [43].

When the reasoning is complete, the *Conversational Explanation Generation* module takes over and conveys the explanation in to the recipient. Conveyance can be an interactive process, using, for instance, a chatbot [36] or voice assistant. This module needs to be aware of the audience (i.e., expert or layperson) and generate conversation accordingly. Recent advances in imbuing conversational agents with personality could address this requirement. Rationalization techniques [13] are also useful here.

These latter two modules are where the bulk of the computational effort and open research questions lie. The systems discussed in [51] (Figure 4) and [34] (Figure 3) implicitly reflect the architecture of an ExplOr.

Generally, the quality of explanations can vary widely, and there is no universal way to compare them against each other. Explanations are sensitive to the models involved, the structure of queries, and the input. We introduce the human binarization component to sidestep this problem and reduce it to a decision problem. In many domains where AI systems augment the capabilities of human operators, this allows the human to reward or penalize the output of an ExplOr.

Another important component of an ExplOr is the *Registry* module, which keeps track of unique inputs, queries, responses, and whether or not those responses were accepted. When identical input/query pairs are received in the future, the ExplOr can instantaneously respond by recalling stored information. Further, because a large number of machine learning models will be involved in query parsing, input parsing, reasoning, and conversational explanation generation, it is important to keep track of the models themselves in a model registry. In other words, the provenance of every explanation is traceable.

5 Discussion

As explainability becomes a non-negotiable requirement in many AI-enabled systems that collaborate with humans, it is increasingly important to be precise about what explanations are and how one is better than the other. We believe the concept of Explanation Oracles can fulfill this need.

ExplOrs decouple the complexity of generating explanations from how they are used. All the systems within the boundary of the ExplOr can grow in complexity or sophistication over time without affecting how they are used. As long as they are within the same domain and present the same interface, humans can use ExplOrs to augment their work. The proposed architecture also allows for the re-use of large parts of an ExplOr's infrastructure for similar use-cases in the same domain.

By explicitly allowing and keeping track of adversarial inputs (when that information is provided), ExplOrs become more robust over time. It is becoming increasingly evident that adversarial input examples are inevitable [37] and detecting these examples has to occur post-classification or outside the classifier itself. [33]. In regulated domains, maintaining a registry of known attacks is important. This is a similar idea to malware protection, where computers routinely receive updates on virus signatures as new threats are discovered in the wild and defenses against them are developed.

Two questions bear reflection here. First, can one ExplOr use other ExplOrs as sources of explanations? There is no reason to preclude this possibility. In complex domains, explanations can take on multiple layers of complexity, and using ExplOrs designed for specific tasks as a scaffold towards the complete explanation is permitted. Interestingly, as long as an ExplOr presents a uniform interface, its internal components can be upgraded transparently. Second, why should the response of an ExplOr be trusted? The most important aspect of an ExplOr is its auditability. Every query and response, along with all the metadata, is recorded for future use. This enables auditability. The track record of the ExplOr is available for inspection, and the human who receives the response has the option of rejecting it if not acceptable. This binarization of the output is a safeguard, but it assumes that the human is not colluding with the ExplOrs.

This architecture also suggests that it could be possible to standardize on specific of ExplOrs in well-defined use cases. When explanations become legally required, it makes sense for multiple entities to use an identical instance of an ExplOr (perhaps even an ExplOr as a service) in order to mitigate the costs and complexities of developing and maintaining them.

Having proposed the concept of an ExplOr in this paper, we will now turn our attention to developing proof-of concept-implementations to understand the practical challenges of implementing them.

Acknowledgements

This draft has benefited greatly from discussions with and suggestions from Andrea Brennen, George Sieniawski, Bob Gleichauf, Zach Lyman, and Vishal Sandesara.

Appendices

A The magic of explanation

A.1 Man behind the curtain: show mechanics

The mechanics of *Penn & Teller: Fool Us* are not quite as simple as the premise. Producers of the show identify and invite magicians to perform on the show. Prior to the trick being performed, the magician has to show how the trick is done [32, 23] to another expert magician who serves as a trusted behind-the-scenes third party. For most of the six seasons, Johnny Thompson [45], aka *The Great Tomsoni* had this role. This is required so there is ground truth on how the trick is done. As the magician is performing the trick on stage, P&T are in constant communication with each other and the Great Tomsoni, so he can ascertain whether or not they have actually figured out the trick. There is one instance [2] where The Great Tomsoni has overruled P&T in favor of the magician. As much as we would all like to believe that the show is based entirely on the

honor system, the trusted third party ensures honest behavior.

There have been 74 regular episodes [46] of the show over its six seasons. 298 acts have attempted to fool P&T with 77 foolers, yielding a successful explanation rate of 74.16%. On average, there are four acts that try to fool P&T per episode and one act succeeds. It is likely that this ratio of approximately 1 in 4 foolers per episode is contrived by the producers for entertainment reasons. While the specific statistics are likely distorted, the important points to take away are: (a) the rate of successful explanation is much better than chance and (b) human experts are satisfied by the quality of the explanations. The statistics are summarized below.

	Number	Percentage
Performers	298	100
Correctly explained	221	74.16
Fooled	77	25.84
Overruled	1	0.34

The most important aspect of each episode is the explanation of the trick itself. It would be a simple matter for P&T to explain the trick in plain language, but this would violate an unwritten rule in the magic community, which largely believes in keeping the *how* of magic tricks hidden from public view. Although this is not a view that P&T completely subscribe [30] to, they explain the trick coded in magic jargon as opposed to plain language. In this way, P&T can demonstrate that they have deciphered the key principles of how the trick was done without giving away its mechanics to the public. There are four typical structures for explanations.

- *Type 1*: You did not fool us. Here is how you did the trick (presented in magic jargon).

- *Type 2*: We are uncertain. This is how we think you did it. If you agree with our reasoning, you did not fool us.
- *Type 3*: You partially fooled us. There is a part of the trick that we were not able to decipher.
- *Type 4*: You completely fooled us. We have no idea how the trick was done.

On the TV show, a Type 3 or 4 explanation results in the performing magician receiving the trophy.

A.2 Four tricks: a glimpse of an ontology of magic

This is the fun part of the paper (wherein you are allowed to watch TV for science). I selected four performances which are amazing magic tricks in their own right but differ markedly in their explanations. Given the setting of the show, magicians don't simply perform well-known tricks but strive to create elements of novelty. In some cases they perform several tricks with a theatrical element connecting them. Each section below represents an instance of one type of explanation. Videos of all the performances are freely available on YouTube so take some time to watch them. [Note: The easiest way to find these videos is to search YouTube for the name of the performer and "Fool Us" or copy and paste the URL included in the references]

A.2.1 Exact Cuts: Jimmy Ichihana (Season 6, Episode 6)

Jimmy Ichihana, who came back to the show for the second time to perform close-up card magic, does two different card routines that involve cutting a deck of cards exactly [19] by color, suit, or number at very high speed. Given the speed at which he presents the effects, some parts of the trick fooled Penn and other parts fooled Teller, but between them they were able to decipher the trick completely and give a *Type 1 explanation.* We can make two observations from this performance.

- *Replaying the input*: Given the rate at which Ichihana was handling the cards and producing effects, it is unlikely that either of P&T were able to decompose the act in real-time and likely had to replay the act mentally to figure out how the trick was done. In magic, the performer has complete control of how the trick proceeds, and can sometimes have multiple paths to the effect, so P&T often will need to replay the trick mentally to decipher it. They have no way of deciphering it on the fly because they don't know what to pay attention to - the whole point of magic tricks is to misdirect viewers and control their attention. ExplOrs will, at a minimum, need to be able to play back the input to do something similar.
- *Two is better than one*: This is a case where having both P&T try to explain the trick was better than having either one of them alone. The

deliberative process after viewing a trick includes generating, evaluating, and discarding or selecting the most viable explanation of the trick. Consider, as a logical extrapolation of this observation, the limiting case of P&T being replaced by every magician who ever lived. In this case the probability of successful fooling is vanishingly small.

A.2.2 Film to life: Rokas Bernatonis (Season 3, Episode 6)

We now look at a Type 2 explanation. Rokas Bernatonis performs [53] a trick that combines two well known tricks called *Card to Wallet* and *Film to Life*. P&T correctly identify the two tricks and then tell Bernatonis in code that they think the combined trick has certain elements and ask Bernatonis if they are wrong. He says they are correct and therefore not fooled.

- *Explanations must quantify uncertainty*: In this case, the explanation of the trick included uncertainty, and this was explicitly mentioned. ExplOrs will need to quantify the uncertainty in the explanations they generate as well.

A.2.3 Composition: Eric Mead (Season 4, Episode 10)

Eric Mead [54] fooled P&T with a close-up coin magic trick that involved sleight of hand. It is worth watching the trick to see how wonderfully it is orchestrated. It felt like there was nothing superfluous in his routine. At the outset, Mead states that the trick he is about to perform is a variant of a Cylinder and Coins trick invented by John Ramsay. It is not clear if this concession is simply part of the "*verbal misdirection, which may or many not have already begun*" intended to force P&T into a mental frame. The look of pure delight on P&T's faces throughout the routine is ample evidence of the mastery demonstrated by Mead. From P&T's commentary at the end of the trick, this was a *Type 3 explanation*.

Two salient points emerge from the dialogue and subsequent explanations

- *Historical knowledge and evolution of tricks*: Magic has a very long history [26, 22] and it is reported that Teller possesses encyclopedic knowledge [40] of magic history. A successful explanation requires recognizing that a specific trick is being performed and knowing whether it has any variants or introduces novelty beyond the original trick. During the performance, Eric Mead indicates this novelty by claiming that "*I'm certain that there are a couple of beats along the way that are puzzling or mystifying to them.*" It would be a lot easier to fool P&T if they did not possess this wealth of historical knowledge and a sense for how it evolves over time. ExplOrs attempting explanations in domains with significant historical depth would need to capture this knowledge and its subsequent drift/evolution in training data.

- *Mutual common knowledge*: Both Mead and P&T have some knowledge of each other's magic repertoire. Mead notes that the trick they are about to see has been performed by P&T "*on this very stage*". Similarly, P&T note that "*we know you and we know this trick*". This is similar to common knowledge in the game theoretic sense [44]. For our discussion, this is equivalent (a) to Mead (at least partially) knowing what P&T know i.e., a gray box or a white box attack and (b) both the ExplOr and the expert having access to common knowledge in the field.

A.2.4 Nothing to go on: Harry Keaton (Season 6, Episode 3)

In this trick, Harry Keaton [52] asks the emcee of the show, Alyson Hannigan, to feel objects hidden inside a minimal rectangular box. When she discloses what she thinks she is feeling, Keaton uncovers the box to reveal a completely different object. For instance, she says she feels a sponge, and Keaton reveals a small rock. P&T had never seen anything like it before - "*. . . we got no way to figure out anything because you invented the damn thing. . . .*". Clearly this is a *Type 4 explanation.* Further research shows Keaton's trick has a long lineage [25] starting in 1954, although P&T are completely unaware of the existence of the effect, and didn't hazard any guesses at explaining it. What can we learn from this explanation?

- *Complete novelty defies explanation in conventional terms*: Whether the novelty is actual or merely perceived because the trick is completely outside the P&T's experience, it is impossible to explain the mechanics of the trick. For an ExplOr an appropriate capability to have in this situation is to be able to recognize that a novel out-of-distribution sample has been encountered and decline to offer explanation.

A.3 Anatomy of an explanation

In this section, we look at the hierarchy of knowledge required to explain a magic trick. Consider Figure 4.

1. *Materials of Magic*: At the base of the hierarchy is a material used to do the trick. These can be playing cards, coins, ropes, elephants, or custom-made props. Materials can be either be in their natural state or *gimmicked* i.e., altered to aid in achieving the effect. Choosing a material narrows down the range of possible explanations in higher levels of the hierarchy.

2. *Misdirection*: Misdirection - focusing the attention of the audience away from the cause of a magic effect - is fundamental to most magic tricks. A notable exception is the class of self-working tricks that depend only on mathematical properties [39, 11]. According to Sharpe [38], misdirection is "*the intentional deflection of attention for the purpose of disguise*". In principle, most magic tricks could be explained by describing how misdirection is used at various stages during the trick. However, this is akin to explaining how swimmers can increase their speed by looking at cellular

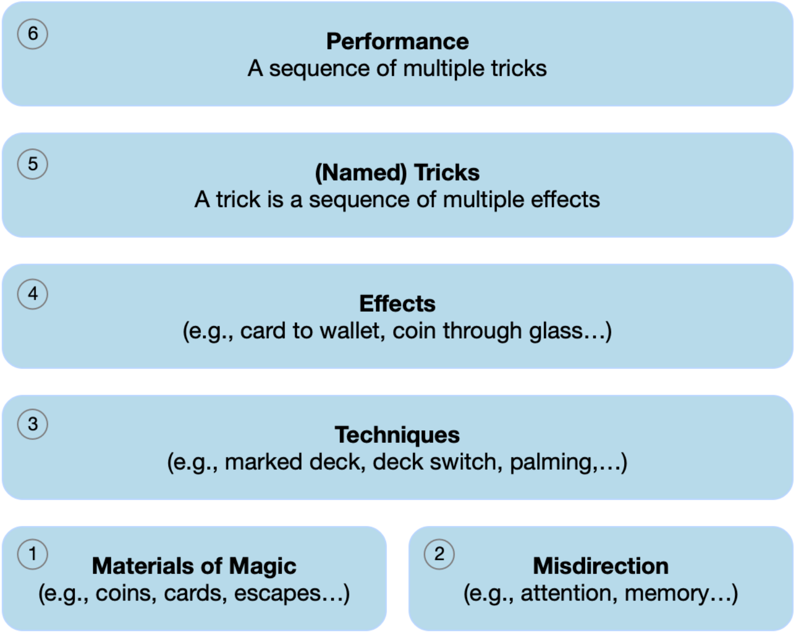
6
Performance
A sequence of multiple tricks
5
(Named) Tricks
A trick is a sequence of multiple effects
4
Effects
(e.g., card to wallet, coin through glass…)
3
Techniques
(e.g., marked deck, deck switch, palming,…)
1
Materials of Magic
(e.g., coins, cards, escapes…)
2
Misdirection
(e.g., attention, memory…)

Figure 4: Representation of the hierarchy of knowledge to explain magic performance.

processes. It is important to tailor the explanation to the audience by selecting an appropriate level of description.

Kuhn et al. [21] propose a psychologically-based taxonomy of misdirection. They connect magic techniques to the psychological principles exploited to achieve them. Over time, magicians have developed techniques to affect our perception, our memory, and our ability to reason about the trick. As a result, their taxonomy has three branches - perception, memory, and reasoning - and about 45 distinct mechanisms of misdirection. While they treat each mechanism as independent, they note that many types of misdirection can be used in parallel.

3. *Techniques*: Think of techniques as the atoms of magic. Magicians use or develop repeatable technique that can be use in many tricks. For instance, The Conjuring Archive [5] catalogs 23,794 card sleights. Each of these is a technique used by magicians to control playing cards. Consider the *double lift*. This is a technique to lift the top two cards from a deck while making it look like only one card was picked up. This move is the basis of many different card tricks.

4. *Effects*: Effects are what the spectators of magic experience. Magicians develop techniques through years of practice but spectators experience effects with no effort. Consider, for example, the striking effect known as

Any card at any number. A spectator calls out any card, say a 4 of spades, and a number, say 13. The magician counts down 13 cards from the top of the deck to reveal the chosen card. There are many techniques that can be use to achieve this effect.

5. *Tricks*: Tricks are collections of one or more effects performed one after the other. Many tricks have names, some after their inventors[1]. P&T's explanations are usually delivered at this level. They only go lower in the hierarchy when there are no named tricks they can use to explain the trick.

6. *Performance*: A performance is a collection of related tricks performed one after another.

Our observations about the explanation hierarchy are listed below. These inform the architecture of ExplOrs in Section 4.

- From Figure 4, *tricks* depend on *effects* which in turn depend on *techniques* which themselves rely on *misdirection.*

- Magicians who create tricks typically start with a material, decide on the effects they want to achieve, and develop techniques to accomplish the goal. In other words *adversaries start at the bottom of the hierarchy and work their way up.* This is similar to adversarial attacks [41] on AI systems.

- Explainers, on the other hand, start at the trick level and migrate to the next level *down* when they deplete their ability to explain at any given level. Their explanations are typically offered at the trick, effect, or technique level. They rarely offer explanations at the misdirection level.

- Explanations are domain-specific **and** audience-specific. In order to be relevant, explanation of an effect based on, for example, misremembering [21] would be different when delivered to a magician and a psychologist. Designers of XAI systems need to recognize this important requirement.

- Multiple techniques can be used to achieve the same effect, and a trick can be composed of effects in multiple ways. This suggests that starting with explanations at the highest possible level is an efficient way to rule out large chunks of the search space. The higher the level the greater the explanatory power.

- Explanations on the show are consistent with the literature on explanations from the social sciences. See, for example, [27, 9]. P&T's explanations are conversational [18] and hew closely to the Gricean maxims [16] of relevance, quality, and quantity. They, however, consciously violate the fourth maxim of avoiding obscurity of expression and ambiguity to protect the secrets of magic.

[1] With the usual controversies about who was the first to invent the trick.

- Finally, we note that there are two distinct stages of explanation on the show. The first is to diagnose how the trick was done and the second is to convey an explanation of that cause in conversational terms to the performing magician. This second phase is often intentionally obfuscated in order to protect the secrets of magic. This two-stage pipeline is in excellent agreement with the conversational model in [18].

References

[1] Alvarez Melis, D., and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7786–7795.

[2] Amodei, I. Penn and Teller most controversial magician who fooled!, 2016. `https://www.youtube.com/watch?v=9WXPwT2kPGo`, Last accessed on 2019-12-30.

[3] Armstrong, S., and O'Rorke, X. Good and safe uses of AI oracles. *CoRR* (2017).

[4] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR* (2014).

[5] Behr, D. Conjuring Archive. https://www.conjuringarchive.com/, Last accessed on 20202-03-28.

[6] Bellare, M., and Rogaway, P. Random oracles are practical. In *Proceedings of the 1st ACM conference on Computer and communications security - CCS '93* (1993).

[7] Bostrom, N. *Superintelligence : Paths, Dangers, Strategies.* Oxford University Press, New York, 2016.

[8] Brennen, A. 10 things I've learned about explainable AI. `https://medium.com/high-stakes-design/10-things-ive-learned-about-explainable-ai-e7f963d6bfc2`, Last accessed on 2019-12-30.

[9] Byrne, R. M. J. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (Macao, China, Aug. 2019), International Joint Conferences on Artificial Intelligence Organization, pp. 6276–6282.

[10] Chen, M., D'Arcy, M., Liu, A., Fernandez, J., and Downey, D. CODAH: an adversarially authored question-answer dataset for common sense. *CoRR* (2019).

[11] Diaconis, P. *Magical Mathematics : The mathematical ideas that animate great magic tricks*. Princeton University Press, Princeton, 2011.

[12] Doshi-Velez, F., and Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. http://arxiv.org/abs/1702.08608.

[13] Ehsan, U., Harrison, B., Chan, L., and Riedl, M. O. Rationalization: a neural machine translation approach to generating natural language explanations. *CoRR* (2017).

[14] Finlayson, S. G., Chung, H. W., Kohane, I. S., and Beam, A. L. Adversarial attacks against medical deep learning systems. *CoRR* (2018).

[15] Gleichauf, B. A chatbot? Are you sirious? `{https://gab41.lab41.org/a-chatbot-are-you-sirious-9a7a615b3cfa},` `Lastaccessedon2020-04-10`.

[16] Grice, H. P. *Syntax and Semantics 3: Speech acts*. Academic Press, New York, 1975, ch. Logic and Conversation, pp. 41–58.

[17] He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)* (10 2017), p. nil.

[18] Hilton, D. J. A Conversational Model of Causal Explanation. *European Review of Social Psychology 2*, 1 (Jan. 1991), 51–81.

[19] Ichihana, J. Jimmy Ichihana makes Exact Cuts. `https://www.youtube.com/watch?v=lWJz1NMT638`, Last accessed on 2019-12-30.

[20] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *CoRR* (2019).

[21] Kuhn, G., Caffaratti, H. A., Teszka, R., and Rensink, R. A. A psychologically-based taxonomy of misdirection. *Frontiers in Psychology 5*, 5 (2014), 1–14.

[22] Lamont, P. *The secret history of magic : the true story of the deceptive art*. TarcherPerigee, New York, 2018.

[23] Lebovitz, D. L. The untold truth of Penn & Teller: Fool Us. `https://www.looper.com/177614/the-untold-truth-of-penn-teller-fool-us/`, Last accessed on 2019-12-30.

[24] LUNDBERG, S., AND LEE, S.-I. A unified approach to interpreting model predictions. *CoRR* (2017).

[25] MAGIC, M. Sensor Box by Boretti. https://www.martinsmagic.com/allmagic/stage/borettis-sensitive-box-by-boretti/, Last accessed on 2019-12-30.

[26] MAGICPEDIA. Timeline of magic. http://geniimagazine.com/wiki/index.php/Timeline_of_Magic, Last accessed on 2019-12-30.

[27] MILLER, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]* (Aug. 2018). arXiv: 1706.07269.

[28] MITTELSTADT, B., RUSSELL, C., AND WACHTER, S. Explaining explanations in ai. *CoRR* (2018).

[29] MOLNAR, C. Interpretable machine learning: A guide to making black box models more interpretable. https://christophm.github.io/interpretable-ml-book/, Last accessed on 2020-04-29.

[30] PANG, K. Penn and Teller are revealing how their magic tricks are done—and it's O.K., 2015. https://www.vanityfair.com/culture/2015/09/penn-and-teller-fool-us-revealing-tricks, Last accessed on 2019-12-30.

[31] PEZESHKPOUR, P., SRINIVASAN, R., AND CHANDER, A. Generating User-friendly Explanations for Loan Denials using GANs. 9.

[32] PIERRO, S. Backstage with Penn & Teller: Fool Us, 2015. https://www.simonpierro.com/penn-teller-fool-us, Last accessed on 2019-12-30.

[33] QIN, Y., FROSST, N., SABOUR, S., RAFFEL, C., COTTRELL, G., AND HINTON, G. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. *CoRR* (2019).

[34] RAUSCHECKER, A. M., RUDIE, J. D., XIE, L., WANG, J., DUONG, M. T., BOTZOLAKIS, E. J., KOVALOVICH, A. M., EGAN, J., COOK, T. C., BRYAN, R. N., NASRALLAH, I. M., MOHAN, S., AND GEE, J. C. Artificial Intelligence System Approaching Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. *Radiology* (Apr. 2020). Publisher: Radiological Society of North America.

[35] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. *CoRR* (2016).

[36] ROLLER, S., DINAN, E., GOYAL, N., JU, D., WILLIAMSON, M., LIU, Y., XU, J., OTT, M., SHUSTER, K., SMITH, E. M., BOUREAU, Y.-L., AND WESTON, J. Recipes for building an open-domain chatbot. *CoRR* (2020).

[37] SHAFAHI, A., HUANG, W. R., STUDER, C., FEIZI, S., AND GOLDSTEIN, T. Are adversarial examples inevitable? *CoRR* (2018).

[38] Sharpe, S. *Conjurers' psychological secrets.* Hades Publications, Calgary, 1988.

[39] Simon, W. *Mathematical magic.* Dover Publications, New York, 1993.

[40] Small, J. 5 reasons you'll want to see Penn & Teller live asap. `https://www.travelzoo.com/blog/see-penn-teller-live-asap/`, Last accessed on 2019-12-30.

[41] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations* (2014).

[42] Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Jin, H., Petryk, S., Bargal, S. A., and Gonzalez, J. E. Nbdt: Neural-backed decision trees. *CoRR* (2020).

[43] Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (- 2019), p. nil.

[44] Wikipedia. Common knowledge (logic). `https://en.wikipedia.org/wiki/Common_knowledge_(logic)`, Last accessed on 2019-12-30.

[45] Wikipedia. Johnny thompson. `https://en.wikipedia.org/wiki/Johnny_Thompson`, Last accessed on 2019-12-30.

[46] Wikipedia. Penn & Teller: Fool Us. `https://en.wikipedia.org/wiki/Penn_%26_Teller:_Fool_Us`, Last accessed on 2019-12-30.

[47] Williams, H., and McOwan, P. W. Magic in the machine: a computational magician's assistant. *Frontiers in Psychology 5* (Nov. 2014).

[48] Williams, H., and McOwan, P. W. Manufacturing Magic and Computational Creativity. *Frontiers in Psychology 7* (June 2016).

[49] Xie, N., Ras, G., Gerven, M. v., and Doran, D. Explainable deep learning: a field guide for the uninitiated. *CoRR* (2020).

[50] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. CLEVRER: Collision events for video representation and reasoning. *CoRR* (2019).

[51] Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenenbaum, J. B. Neural-Symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems* (2018), pp. 1039–1050.

[52] YouTube. Magician fooled Penn & Teller without moving!! `https://www.youtube.com/watch?v=eAKUajF0LYE`, Last accessed on 2019-12-30.

[53] YouTube. Penn & Teller Fool Us Rokas Bernatonis. https://www.youtube.com/watch?v=TTgyIqx0ZDM, Last accessed on 2019-12-30.

[54] Youtube. Penn & Teller: Fool Us Eric Mead - What the coins?, 2017. https://www.youtube.com/watch?v=9w1vMWsB3wc, Last accessed on 2019-12-30.

[55] Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. SWAG: a large-scale adversarial dataset for grounded commonsense inference. *CoRR* (2018).