



uOttawa

CSI5101 Knowledge Representation

Winter 2024

Assignment 4 Trust in KA and Neuro-Symbolic AI

Submitted by

Romario Vaz : 300308477

Saranya Krishnasami : 300321456

Index

Trust in KA and NeuroSymbolic AI

Index	2
Trust in KA and NeuroSymbolic AI	2
Q1 Trust in Knowledge Acquisition Q1 – Sources for KA (6 points)	3
Q2– EAT (5 points)	6
Q3 – Fake news (8 points)	9
Q4 – InterAnnotator Agreement (5 points)	13
Q5– Recommendation systems (4 points)	15
Q6 – Fallacies (4 points)	16
Q7 – 3 waves of AI (9 points)	18
Q8 – Models interpretability (5 points)	21
Q9 – Explainability (8 points)	24
Q10 – Future of NeuroSymbolic AI (6 points)	27

Q1 Trust in Knowledge Acquisition Q1 – Sources for KA (6 points)

a. Tradition:

Growing up, I learned the specific method of rolling dough on a fork to produce a *kulkul*, which is a traditional Christmas snack from India (Sarkar, 2022). Every Christmas, it was a family tradition to watch a movie while rolling the dough balls, which is how this type of knowledge was acquired.

b. Borrowing:

One of my former coworkers is extremely proficient with Excel shortcuts and is able to work productively while only using the keyboard and rarely ever touching his mouse. By sitting next to him, I picked up these skills and "borrowed" this piece of knowledge.

c. Role Modelling & Mentorship:

When I was growing up, I noticed my mother always investing her money every time her salary came in. Every time I wanted her to buy me something, she would tell me that she needs to invest the money into assets rather than buying liabilities like toys, video games, etc. Learning from her as a role model and mentor, when I first started earning my own money, I ensured that I invested in wealth producing assets first, before spending money on discretionary expenses.

d. Logical Reasoning:

There are many public transport options from my house to the university, including at least 4 buses and the OTrain. After living here for 2 years and understanding the bus schedules, reliability, delays, etc., I can use logical reasoning to determine the fastest route to get to the university depending on what time I leave the house.

e. Scientific Approach:

The scientific approach typically involves asking a question, performing background research, hypothesizing, experiments, and forming a conclusion based on the results and analysis (Wright & Lavery, 2023). Last year I noticed that I was spending many hours a day

on social media, and during the same time, I found it hard to keep up with my studies. As a result, I asked myself if my social media use was causing me to struggle with school. I did some research on productivity, and was reinforced with the idea that eliminating distractions is a great way to become more productive. Hypothesizing that gradually reducing social media usage would improve productivity, I began tracking the amount of time spent on social media per day on a spreadsheet. Over the course of a year, I had reduced my usage from an average of 1.72 hours a day to 0.54 hours per day of social media usage. In this time, my GPA increased by one grade point and I also found myself less exhausted with my university workload. Looking at the results, I concluded that since I was able to reduce social media usage, I was able to both increase my GPA and reduce how overwhelming my studies felt, which was a successful implementation of the scientific method.

f. Intuition:

My friend loves Harry Potter. He has read all the books and watched all the movies several times. Based on my familiarity with his tastes, I was able to use my intuition to acquire the knowledge of the kind of birthday gift he would like. Therefore, when we travelled together to London, I bought him the perfect birthday gift of an exclusive Warner Brothers Studio Tour of the Harry Potter filming set.

g. Trial & Error:

I acquired the knowledge of how much parmesan goes with spaghetti bolognese through trial and error. When I first started cooking it, I would add too much parmesan and find that it tasted too salty and the cheese would overpower the taste of the meat and tomatoes. Sometimes I would add too little and would find that the dish lacked flavour. After several attempts at cooking the dish over time, I figured out the correct ratio to use.

h. Experience:

After studying computer science for over 6 years, working on group projects taught me the importance of version control systems like Git. Initially, managing changes and collaborating with others on code was chaotic because we would just keep renaming files with suffixes like "_Final", "_FinalVersion2", "_FinalVersionREAL", etc. With experience, I

learned to effectively use branches, merge requests, and commits to keep the project organized and everyone on the same page.

i. Authority

When a doctor prescribes medication or a treatment plan, I follow it based on their authority and expertise in medical science. Their years of study and practice provide knowledge that I trust for my health decisions.

References:

Sarkar, P. V. (2022, November 15). Make kulkuls, Indian sweet curls, for Christmas. The Spruce Eats. <https://www.thespruceeats.com/kulkulsindianchristmassweetcurls1957799>

Wright, G., & Lavery, T. (2023, February 10). What is the scientific method and how does it work?: Definition from TechTarget. WhatIs. <https://www.techtarget.com/whatis/definition/scientificmethod>

Q2- EAT (5 points)

The 2 selected articles are "How to rank higher on Google in 13 steps from Backlinko" (Dean, 2024) and "How to Rank Higher on Google in 2024: 15 SEO Tips To Conquer Search" (Molenaar, 2023).

Five main recommendations from these articles and their relationship with Google's EAT (Google, 2024) are:

i. Request backlinks to the site (Dean, 2024): The idea is to get other websites to link to the site by emailing them with helpful information and requesting to link to the site. This helps build authoritativeness (A) by getting other websites to link back to the website.

ii. Publish high quality content (Dean, 2024): If an article is perceived as higher quality, more people are expected to share it on social media and add links to the article. This builds authoritativeness (A) by having an article be a data source of valuable information that others link to frequently.

iii. Technical SEO (Molenaar, 2023): Ensure that the pages are secure and optimized. It is essential that HTTPS is used rather than HTTP to ensure that user data (especially sensitive information like payments) is secured. This helps build the site's Trustworthiness (T).


iv. Hire Experts (Molenaar, 2023): In order to improve the content of the site (as well as design and other noncontent components), one way to improve search rankings is to hire someone with specialized knowledge in the field of interest. This helps build the site's Expertise (E).

v. Create a Google My Business Profile (Molenaar, 2023): Using Google My Business helps impact local search results and provides information like business hours, contact details, and address, and also allows users to leave reviews. This helps boost the site's Trustworthiness (T).

Example of site where these recommendations are followed:

healthline [Health Conditions](#) [Discover](#) [Plan](#) [Connect](#) [SUBSCRIBE](#) [Q](#)

Everything You Need to Know About Kidney Stones

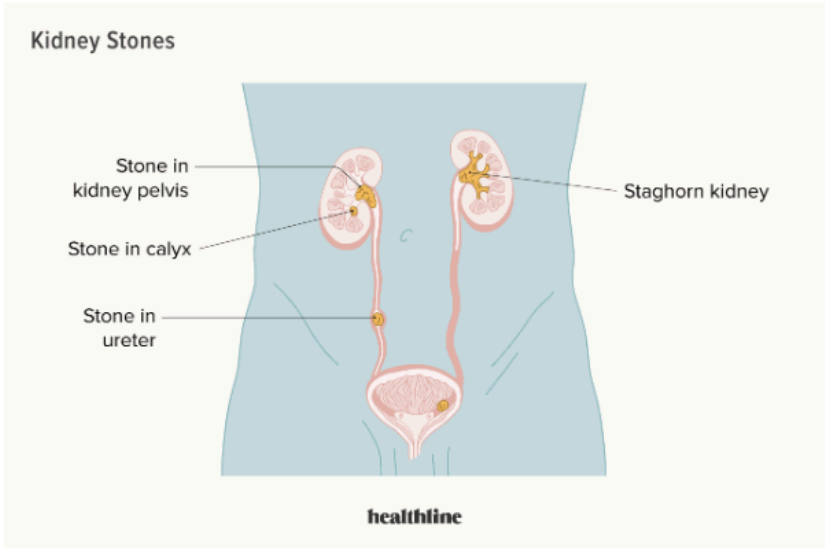


Medically reviewed by [Alana Biggers, M.D., MPH](#) — By [The Healthline Editorial Team](#) — Updated on [February 6, 2023](#)

[Symptoms](#) | [Causes](#) | [Treatment](#) | [Diagnosis](#) | [Passing a kidney stone](#) | [Prevention](#) | [Foods to avoid](#) | [When to see a doctor](#) | [Takeaway](#)

Kidney stones, or renal calculi, are solid masses made of crystals. They can develop anywhere along your urinary tract, which consists of the kidneys, ureters, bladder, and urethra.

Kidney stones can be a painful medical issue. The causes of kidney stones vary according to the type of stones.



healthline

Not all kidney stones are made up of the same crystals. The different types of kidney stones

The Healthline site tends to follow these recommendations, as can be seen in the above example (Healthline Media, 2023). Healthline is often cited (i.e., "backlinked") by other reputable websites in the health and wellness space, as well as in academic papers and studies, due to its comprehensive and authoritative content. The quality of articles is typically high with lots of visuals and easy to understand explanations. The site is secure (using HTTPS), fast, mobile friendly, and has a clear structure that makes it easy for both users and search engines to navigate. Healthline works with medical professionals,

specialists, and experts to review and contribute to their content (the above article is medically reviewed by a certified MD). To summarize, points iiv listed above are followed well (point v is not followed because Healthline is an online business rather than a local business, and hence, does not use Google My Business).

References:

Dean, B. (2024, March 14). How to rank higher on Google in 13 steps [2024]. Backlinko.
<https://backlinko.com/rankhighongoogle>

Molenaar, K. (2023, October 27). How to rank higher on Google in 2024: 15 seo tips to conquer search. Influencer Marketing Hub.
<https://influencermarketinghub.com/howtorankhigherongoogle/>

Google. (2024, March 5). Google quality rater guidelines googleusercontent.com.
<https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf>

Healthline Media. (2023, February 6). Kidney stones. Healthline.
<https://www.healthline.com/health/kidneystones>

Q3 – Fake news (8 points)**A. Easy Feature: Check the Date**

Date extraction is straightforward with many date parsing libraries. Typically dates are in a standard format near the beginning of an article within appropriate HTML tags.

Simple algorithm:

Extract metadata and header text of input news article

Use a date parsing library to extract the date from the extracted text.

Check if the extracted date is older than a given threshold.

This algorithm should work in most cases. Some cases it could fail are when an article does not have a date mentioned, or where it is in a nonstandard format.

B. Moderate Feature: Check Mainstream Media

The title could be extracted and searched through several mainstream media sites. If there is a high semantic textual similarity with the title in many different sites, then it is likely that the same story has been published by other sites. This is moderately difficult because automating the process of searching a news site with a text is nontrivial. Some sites provide an API, while others require automated crawling tools like Selenium to create a search query. Further, textual semantic similarity is computationally expensive if using a large language model like BERT or GPT, and typically requires a GPU. Simpler embeddings like TFIDF are less computationally expensive but typically less accurate.

Simple algorithm:

Input: News article content N , List of mainstream media sites M , Similarity threshold θ

Output: Count of mainstream articles with similarity above θ

procedure CHECK_MAINSTREAM_MEDIA(N, M, θ)

 Initialize count $C \leftarrow 0$

 Extract title T from N

```
Get text embeddings  $T_N$  of  $N$ 

for each site  $s \in M$  do

    Retrieve top 3 articles  $A_1, A_2, A_3$  by querying  $s$  with the title  $T$ 

    for each article  $a \in \{A_1, A_2, A_3\}$  do

        Get text embeddings  $T_a$  for article  $a$ 

        Compute cosine similarity  $\sigma$  between  $T_N$  and  $T_a$ 

        if  $\sigma > \theta$  then

             $C \leftarrow C + 1$ 

        end if

    end for

end for

return  $C$ 
```

Some cases that this algorithm would not work include sites with paywalls and access restrictions, very recent news that might not yet have been reported by mainstream media, lower cosine similarity if different media sites have different biases and framing, lower textual similarity when comparing with mainstream media sites that use video content over text, and so on. However, it should still be able to handle most cases reliably well if the set M is well defined and if the search of the sites in the set is performed correctly.

C. Difficult Feature: Is it a joke?

Automated humour detection is a difficult task for several reasons including the inherent subjectivity of humour. Even powerful large language models (LLMs) like GPT4 struggle with detecting humour (Góes et al., 2023).

A simple algorithm:

Finetune a pre trained large language model with a binary classification task of detecting if a piece of text is a joke or not.

Pass the input news article through this model and use the output for the fake news detection task

This algorithm would work on cases where the underlying joke detection model works well. It would most likely struggle in cases where the humour is more subtle, especially in cases where the language and presentation may mimic a genuine news format without obvious humor cues.

Two features not mentioned in the slide:

a. Does the article contain Deepfakes?: AI-generated images and videos of real people (known as Deepfakes) performing questionable actions could be used by malicious actors who create fake news. For instance, there was a viral video of former US president Obama disparaging former US president Trump (BuzzFeed, 2018), which was revealed to be a deepfake within the video (created for informative and entertainment purposes). Commercial deepfake detection algorithms are fairly accurate (Simonchik, 2024) and the presence of deep fakes in a potential fake news article could be determined with such an algorithm, and thus be used as a feature for fake news detection.

b. Writing Quality: Highly reliable news sites follow strict guidelines on their journalistic diction. The presence of poor writing quality, grammatical errors, spelling mistakes, etc. could easily be determined by a variety of text analysis algorithms, and this could then be used as a feature for fake news detection.

References

Góes, F., Sawicki, P., Grze's, M., Brown, D., & Volpe, M. (2023) Is GPT4 Good Enough to Evaluate Jokes? https://computationalcreativity.net/iccc23/papers/ICCC2023_paper_89.pdf

BuzzFeed Video, (2018, April 17). You won't believe what Obama says in this video! YouTube. <https://www.youtube.com/watch?v=cQ54GDm1eL0>

Simonchik, K. (2024, February 21). Deepfake detection: Accuracy of Commercial Tools. LinkedIn.

https://www.linkedin.com/pulse/deepfakedetectionaccuracycommercialtoolskonstantinsimonchiku0z3e/?trk=articlelessrfrontendpulse_morearticles_relatedcontentcard

Q4 – InterAnnotator Agreement (5 points)**5 tasks in order from low kappa to high kappa**

- a. Selecting which flavour of ice cream tastes better: Taste is highly subjective, and most people like ice cream. All flavours are usually sugary and trigger pleasurable feelings within the brain. Many people have a flavour preference for which they cannot justify why their selected flavour is better than someone else's.
- b. Selecting which artwork is better: Art is subjective because appreciation of art varies greatly between individuals due to personal tastes, cultural backgrounds, and emotional responses. However, some pieces are more impressive from a technical perspective, which could influence rankings despite differing tastes among people. For example, almost everyone would argue that the Mona Lisa by Leonardo da Vinci is more beautiful than a stick figure drawn by a 4year old.
- c. Selecting the more attractive person from images of two people: Beauty is said to be in the eye of the beholder and everyone has their preferences, but there are many features that society as a whole considers more conventionally attractive, even among different cultures. For example, facial symmetry, youthfulness, etc. are highly valued in nearly all cultures (Voegeli et al., 2021).
- d. Classifying animal species from an image For a layperson, it is quite easy to identify an animal species from an image. The only possible disagreements are when distinguishing between two highly similar species, like a crocodile and an alligator, or perhaps two similar types of bacteria. For more biologically familiar annotators, there is even less subjectivity.
- e. Solving simple arithmetic There is no subjectivity involved in answering questions like "Which of these two options is equal to 12 minus 5?" and everyone with a basic elementary school education should reasonably be expected to come up with the exact same answer.

Three subtasks that could lead to more agreement about movie appreciation:

- a. Technical Quality

"How would you rate the technical quality of the movie, including cinematography, sound design, and special effects, on a scale from 1 to 10?" This question focuses on the technical aspects of the film, which are generally more objective and easier to evaluate. Reviewers are more likely to agree on whether the cinematography was visually appealing or if the sound design was immersive.

b. Performance by the Cast

"On a scale from 1 to 10, how would you rate the performances of the main cast members?" While there is a subjective element to evaluating acting, focusing specifically on whether the actors convincingly portrayed their characters or evoked the intended emotions can lead to a higher level of agreement among reviewers.

c. Storyline and Script: "How would you rate the storyline and script of the movie for coherence, originality, and engagement on a scale from 1 to 10?" Focusing on the script's coherence (how well the story holds together), originality, and its ability to engage the audience provides a more structured way to assess a film's narrative aspects

To synthesize these subtasks into an overall movie quality evaluation that is less subjective, one could average the scores from the three categories to obtain a composite score that reflects the film's overall quality. Alternatively, one might weigh these categories differently based on the genre of the movie or the preferences of the audience (e.g., weighting technical quality higher for a sci-fi movie than for a drama). The combined score could then be categorized into broader quality bands (e.g., Excellent, Good, Fair, Poor) to simplify the final evaluation.

References:

Voegeli R, Schoop R, PrestatMarquis E, Rawlings AV, Shackelford TK, et al. (2021) Crosscultural perception of female facial appearance: A multiethnic and multicentre study. PLOS ONE 16(1): e0245998. <https://doi.org/10.1371/journal.pone.0245998>

Q5- Recommendation systems (4 points)

When I open Netflix, the first thing I see below "Continue Watching" is a list of currently trending shows. These shows are being suggested to me, seemingly regardless of my viewing history. Below that I see "Suspenseful TV Shows" with a list of suspenseful TV shows and "Because you watched Archer" with a list of shows similar to the show Archer, which I have been watching a lot recently. To summarize, Netflix is primarily pushing trending shows into my main viewing focus on the screen, and only when I scroll down do I see shows that are similar to what I have been watching, i.e., "Suspenseful TV shows" and "Because you watched Archer".

Regarding how easy or hard it is to filter content and have a more diverse selection, on Netflix, there is an option to browse by genres, but the subcategories do not seem to follow any systematic pattern, and appear to be like a folksonomy. Further, the recommendations are still heavily influenced by past activity, making it somewhat challenging to break out of one's content bubble. Discovering completely new content that's outside of what the algorithm thinks one likes requires a bit of effort and deliberate searching.

I feel like the practice of pushing specific content towards users can be a double edged sword. On one hand, it helps in discovering shows or movies that align with my interests, which I might not have found otherwise. It makes the browsing experience more tailored and, arguably, more efficient. I appreciate not having to sift through endless options to find something that matches my tastes. However, I don't think this practice is without its drawbacks. It can create a sort of echo chamber, limiting exposure to diverse content and perspectives. This can be particularly limiting when it comes to political or educational content, where exposure to a wide range of ideas and stories is beneficial. It feels like the balance between personalization and discovery is skewed too much towards the former.

Q6 – Fallacies (4 points)

1. Straw Man Fallacy

Description: The straw man fallacy occurs when someone misrepresents an opponent's position or argument to make it easier to attack or refute. The person sets up a "straw man" version of the argument that is distorted or exaggerated, then proceeds to knock it down, claiming to have refuted the original argument.

Example:

Person A: "I think we should increase funding for public education."

Person B: "You want to waste taxpayer money on an inefficient and bloated education system that doesn't work. I can't agree with that."

Why it's convincing: The straw man fallacy is convincing because it allows the person to avoid addressing the actual argument being made. By creating a distorted version of the argument, they can more easily attack it and make their own position seem stronger in comparison, even though they haven't actually refuted the original argument. This can sway people who are not closely examining the logic.

2. Gambler's Fallacy

Description: The gambler's fallacy occurs when someone believes that the probability of a random event is influenced by previous outcomes, even though each event is independent. For example, believing that if a coin has landed on heads several times in a row, it is more likely to land on tails next.

Example: "The roulette wheel has landed on black 5 times in a row, so it must be due to land on red next."

Why it's convincing: The gambler's fallacy is convincing because it plays on our natural tendency to seek patterns and expect randomness to "even out" over time. Even though each coin flip or roulette spin is independent, the human mind has difficulty accepting that truly random events can produce streaks. This leads people to incorrectly believe that past results can influence future probabilities.

3. Anecdotal Fallacy

Description: The anecdotal fallacy occurs when someone uses personal experience or an isolated example instead of sound evidence as the basis for a general conclusion.

Example: "I know organic food is healthier because my neighbor lost 20 pounds after switching to an all organic diet."

Why it's convincing: The anecdotal fallacy is convincing because personal stories and examples feel more relatable and compelling than abstract statistics or research. People tend to give more weight to individual experiences, even though a single anecdote does not provide sufficient evidence to support a broad generalization. This fallacy plays on our natural tendency to find personal narratives persuasive.

4. The Texas Sharpshooter Fallacy

Description: The Texas Sharpshooter Fallacy is named after a Texan anecdote where a shooter fires at a barn wall and then paints a target around the bullet holes, claiming precision. This fallacy occurs when someone cherry-picks data clusters to support a predetermined conclusion, ignoring evidence that contradicts it or suggests the clusters aren't statistically significant.

Example: A pharmaceutical company conducts a study on a new drug and finds that a subgroup of patients experienced significant improvements in their symptoms. They then highlight only this subgroup's results, omitting data from other groups that didn't show improvement, to claim the drug's effectiveness.

Why it's convincing: The Texas Sharpshooter Fallacy can be persuasive because it presents a seemingly coherent pattern or correlation supporting the argument or conclusion. By selectively focusing on specific data clusters, it creates the illusion of a strong case. However, without considering the full spectrum of evidence, including data that may challenge the conclusion, the argument lacks validity and may mislead others.

References:

HubSpot Blog. (2022, July 26). 16 Common Logical Fallacies and How to Spot Them. Retrieved from <https://blog.hubspot.com/marketing/common-logical-fallacies>

Farmer GD, Warren PA, Hahn U. Who "believes" in the Gambler's Fallacy and why? J Exp Psychol Gen. 2017 Jan;146(1):6376. doi: 10.1037/xge0000245. PMID: 28054813; PMCID: PMC5215234.

Logical Fallacies – Definition and Fallacy Examples. Retrieved from <https://www.freecodecamp.org/news/logicalfallaciesdefinitionfallacyexamples/>

Q7 – 3 waves of AI (9 points)

Analyzing the Similarity based Reasoning question through the Lens of the Three Waves of AI.

First Wave (Handcrafted Knowledge)

In the first wave of AI, the system would rely on a set of predefined rules and knowledge to determine which words are alike and which one is different. The engineer would need to explicitly program the system with information about the properties of the words, such as whether they are chemical elements, professions, or something else. The system would then apply logical rules to compare the words and identify the odd one out.

Second Wave (Statistical Learning)

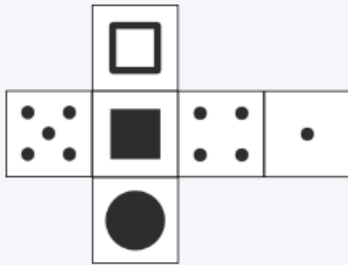
A second wave AI system would approach this problem by analyzing patterns in large datasets of words and their relationships. The system could be trained on a corpus of text data to learn associations between words, their semantic meanings, and how they group together. By applying statistical models, the system could identify the three words that are most similar to each other based on their cooccurrence and contextual usage, and then select the fourth word that does not fit that pattern.

Third Wave (Contextual Adaptation)

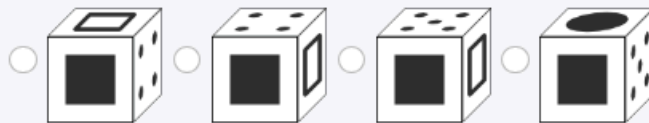
A third wave AI system would go beyond just pattern recognition and statistical modeling. It would construct its own explanatory models of how words and concepts are related, drawing upon broader knowledge about the world. The system would use logical reasoning to understand the underlying principles that make the three words alike, such as their shared properties as chemical elements or professions. It could then apply this contextual understanding to identify the word that does not fit that conceptual framework.

Analyzing a Different Test: Spatial Reasoning

2.



Which cube cannot be made based on the unfolded cube?



First Wave (Handcrafted Knowledge)

For a spatial reasoning test, a first wave AI system would rely on predefined rules and algorithms for tasks like mental rotation, shape recognition, and spatial visualization. The engineer would need to program the system with explicit knowledge about geometric shapes, their properties, and the transformations they can undergo.

Second Wave (Statistical Learning)

A second wave AI system would approach spatial reasoning by learning from large datasets of spatial information, such as images and diagrams. The system could use machine learning techniques to identify patterns and statistical relationships between shapes, their orientations, and how they are transformed. This would allow the system to make predictions and classifications about spatial tasks, even if it doesn't have a deep understanding of the underlying principles.

Third Wave (Contextual Adaptation)

A third wave AI system for spatial reasoning would construct its own models of how spatial concepts and relationships work. It would develop an intuitive, contextual understanding of

geometry, physics, and the mental processes involved in spatial tasks. The system could then apply this knowledge to reason about spatial problems, explain its thought process, and adapt its approach to novel situations.

Analyzing a Game: Chess

First Wave (Handcrafted Knowledge)

In the first wave, a chess playing AI system would be programmed with a set of rules about the movement of chess pieces, as well as strategies and tactics for gameplay. The engineer would need to explicitly define the system's knowledge about the game, including the legal moves, board positions, and common opening/closing sequences.

Second Wave (Statistical Learning)

A second wave chess AI would learn from a large database of chess games to identify patterns and successful moves. It could use machine learning techniques to analyze the statistical probabilities of different board positions and sequences of moves, and then use that knowledge to make predictions and decisions during gameplay.

Third Wave (Contextual Adaptation)

A third wave chess AI would go beyond just pattern recognition and move prediction. It would develop a deeper, contextual understanding of the game of chess, including the strategic concepts, the psychology of the players, and the broader context of the game. This would allow the system to reason about the game, adapt its strategies, and even explain its thought process and decision making to human players.

References:

ixambee. (n.d.). Reasoning Aptitude Previous Year Papers. Retrieved from <https://www.ixambee.com/questions/reasoningaptitude/previousyearpapers/305708>.

JobTestPrep. (n.d.). Pipefitter Practice Test. Retrieved from <https://www.jobtestprep.com/pipefitterpracticetest>.

Q8 – Models interpretability (5 points)

Article 1:

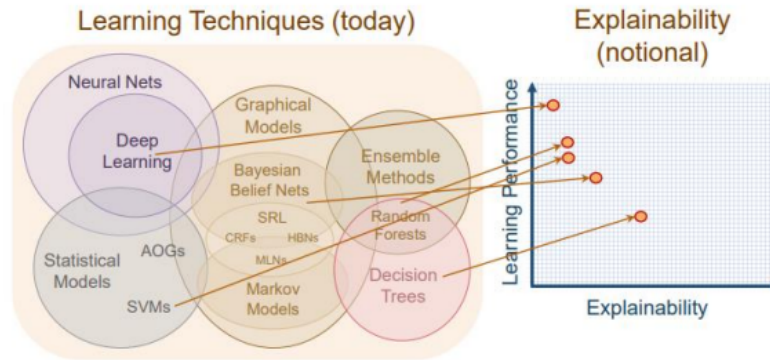


Figure 2: Relationship between explainability and performance of a model (DARPA 2017) [source](#)

Focus: Article 1 (Guest Blog, 2024) primarily focuses on Explainable AI (XAI), its importance, techniques for interpretation, and the tradeoff between accuracy and interpretability.

Key Points:

- Emphasizes the importance of XAI in building trust, ensuring transparency, and enabling accountability in AI systems.
- Discusses techniques for model interpretation, including traditional methods like exploratory data analysis and modern techniques like LIME, SHAP, ELI5, and SKATER libraries.
- Highlights the tradeoff between accuracy and interpretability, where stakeholders may prefer more interpretable models like linear regression and decision trees despite their lower accuracy.

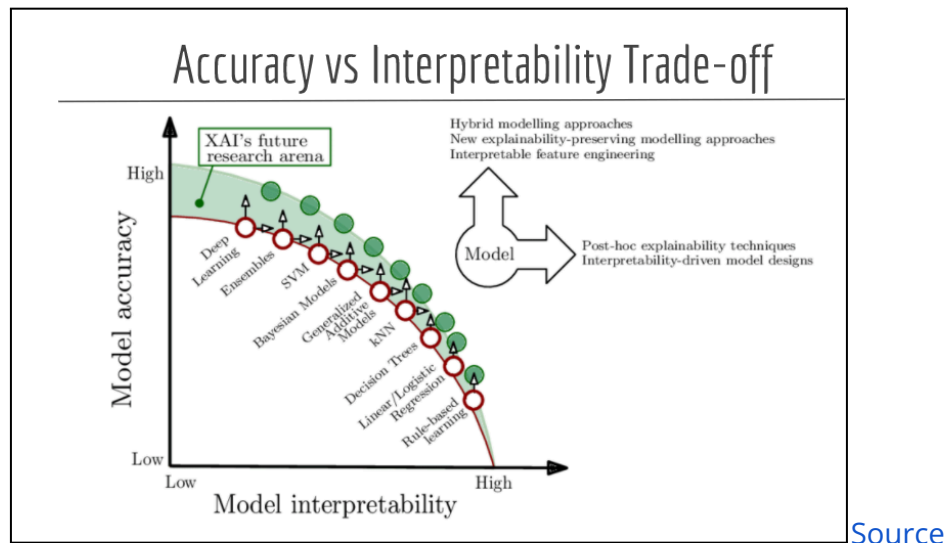
Article 2:

Focus: Article 2 (Tran, N. 2020) focuses on defining explanations, characteristics of explanations, and the evaluation of interpretability. It also discusses interpretability in machine learning and the concept of transparency.

Key Points:

- Defines explanations as making something understandable for others and highlights characteristics of explanations such as accuracy, fidelity, consistency, stability, and comprehensibility.
- Discusses interpretability as a tool for experts to make decisions about models, emphasizing the importance of transparency.

- Proposes three levels of transparency in AI systems: implementation, specifications, and interpretability.



Comparison:

Overlap: Both articles emphasize the importance of understanding AI model decisions and the need for transparency in AI systems. They discuss techniques for achieving interpretability, although Article 1 focuses more on model interpretation techniques, while Article 2 discusses characteristics and evaluation of explanations.

Agreement in Categorization: Both articles categorize techniques and characteristics of explanations and interpretability, in slightly different ways. While Article 1 categorizes techniques based on their application (global interpretation, local interpretation, model agnostic techniques). For instance, article 1 provides an example of global interpretation by discussing how a neural network model for predicting house prices might prioritize the number of squared feet as an important feature, offering insight into the model's decision-making process at a broader level. Article 2 categorizes characteristics of explanations (expressive power, translucency, portability, etc.) and levels of transparency in AI systems (implementation, specifications, interpretability). Article 2 discusses the importance of transparency in achieving interpretability, highlighting how users' ability to understand a model's internal mechanisms can impact its performance. It emphasizes that while fully transparent models may be unrealistic, different levels of transparency contribute to interpretability, thereby affecting performance.

Axis of Performance and Interpretability: Both articles implicitly acknowledge the axis of performance and interpretability. Article 1 discusses the tradeoff between accuracy and interpretability, implying that more interpretable models may sacrifice some accuracy. Article 2 also touches on the importance of understanding the model's internal mechanisms for achieving interpretability, which can impact its performance.

Conclusion:

While both articles approach the topic of explainability and interpretability from different angles, there is an agreement in their categorization alongside the axis of performance and interpretability. Both recognize the importance of understanding model decisions, the tradeoff between accuracy and interpretability, and the need for transparency to achieve interpretability in AI systems. They categorize techniques and characteristics of explanations and interpretability, albeit in slightly different ways, reflecting the multifaceted nature of this complex topic.

References:

- Guest_Blog. (2024, February 06). Explain How Your Model Works Using Explainable AI. Analytics Vidhya.
<https://www.analyticsvidhya.com/blog/2021/01/explain-how-your-model-works-using-explainable-ai/>
- Tran, N. (2020). Towards Explainability in Machine Learning: a User-Centered Perspective. Hamburg University of Applied Sciences.
<https://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/bachelor/tran.pdf>

Q9 – Explainability (8 points)**Summary of Local Model-Agnostic Methods:**

(Molnar, 2023) Interpretable machine learning book is a guide to make black box models explainable to suit business context. Chapter 9.2 discusses using surrogate models to explain predictions of black box models. LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, 2016) explains the individual predictions of black box machine learning models. It utilizes local surrogate models to approximate the predictions of the underlying model for a specific instance of interest. LIME works by perturbing the dataset, generating new samples, and training interpretable models on these variations to explain the predictions effectively. While LIME offers advantages such as flexibility across different types of data (tabular, text, images) and the ability to use interpretable features for explanations, it also faces challenges such as defining a meaningful neighborhood, sampling issues, instability of explanations, and susceptibility to manipulation.

For **text data**, LIME generates variations by randomly removing words from the original text and represents the dataset with binary features for each word, where a feature is 1 if the corresponding word is included and 0 if it has been removed.

For **tabular data**, LIME creates new samples by perturbing each feature individually, drawing from a normal distribution with mean and standard deviation taken from the feature.

When it comes to **images**, LIME segments the image into "superpixels" and turns superpixels off or on to generate variations, allowing for explanations based on the presence or absence of these superpixels in the image.

Discussion from additional source:

The article (Giorgia, 2020) LIME: explain Machine Learning predictions article discusses the LIME (Local Interpretable Model-agnostic Explanations) algorithm, focusing on its steps, intuition, challenges, and advancements. It provides a detailed explanation of how LIME works, including its model-agnostic nature, local explanation focus, and the steps involved in generating explanations. **Additionally**, it delves into the intuition behind LIME, like how it finds the tangent of a complex prediction curve at a specific point. The article also highlights challenges faced by LIME, such as the generation step and selecting the correct kernel width for weighting points, and mentions recent advancements like OptiLIME for addressing these issues. **OptiLIME** provides freedom to choose the best adherence-stability trade-off level and more importantly, it clearly highlights the mathematical properties of the retrieved explanation

The software package that implements LIME (Local Interpretable Model-agnostic Explanations) is the LIME-python (2021) package. The **LIME Python package** is an

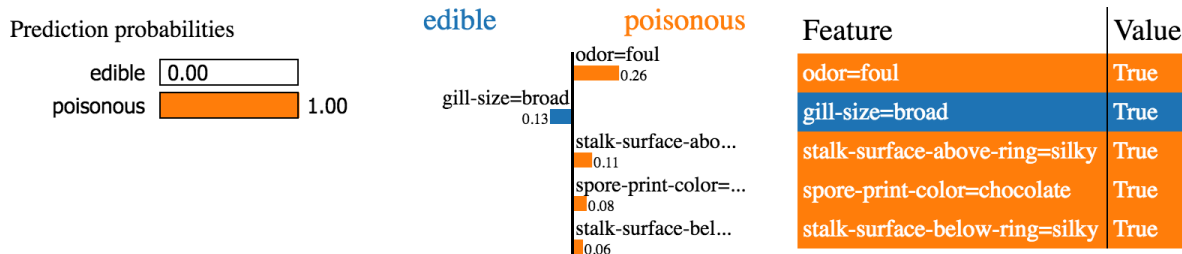
open-source library that provides a way to explain the predictions of any machine learning classifier in an interpretable and faithful manner. The key features of LIME are:

1. Degree of faithfulness: LIME aims to explain the predictions of a model locally around a specific instance, rather than globally. This means it tries to approximate the complex black-box model with a simpler, more interpretable model in the vicinity of the prediction being explained.
2. Model-Agnostic: LIME is designed to work with any machine learning model, regardless of the underlying algorithm or architecture. It does not require access to the model's internals, making it applicable to a wide range of models.
3. Interpretability: LIME generates explanations that are interpretable to humans, such as highlighting the most important features that contributed to a prediction.

LIME applies to both text and image domains, allowing users to interpret the predictions of models in these areas. For example, in the image domain, LIME can highlight the most important pixels that contributed to a classification decision.



LIME Example - Image 1 (explaining prediction of 'cat' [source](#)). Image 2 explaining the image classification prediction.



LIME Example - Mushroom dataset [source](#)

Some examples of using LIME for text and image model interpretability are given above. In summary, the LIME Python package is an open-source software implementation of the

LIME framework, which provides a way to explain the predictions of any machine learning model in an interpretable and faithful manner.

References:

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

Josua .N (2022), How to Interpret Black Box Models using LIME (Local Interpretable Model-Agnostic Explanations).

<https://www.freecodecamp.org/news/interpret-black-box-model-using-lime/>

LIME-python (2021), <https://pypi.org/project/lime-python/>

Q10 – Future of NeuroSymbolic AI (6 points)

The two chosen terms for the NeuroSymbolic AI are Symbolic Neuro symbolic and Neuro-> Symbolic.

Symbolic Neurosymbolic:

The Symbolic Neurosymbolic approach combines symbolic reasoning techniques with neural networks. Symbolic reasoning involves using logic, rules, and explicit representations to manipulate symbols, while neural networks excel at learning patterns from data in a more implicit, subsymbolic manner.

An example of Symbolic Neuro symbolic integration could be in the medical domain, where a system needs to reason about complex diseases and treatments. The symbolic component could encode medical knowledge as logical rules and ontologies, allowing the system to apply logical inference and reasoning. The neural component could then be used to learn patterns from patient data, medical images, and other unstructured information to complement the symbolic knowledge. This hybrid approach could enable the system to both apply expert medical knowledge and learn from data, leading to more accurate diagnoses and treatment recommendations (Anticevic, 2014).

Neuro -> Symbolic:

The Neuro -> Symbolic approach refers to a transition or conversion from neural representations to symbolic representations. This could involve taking the learned features or knowledge from a neural network and translating them into a symbolic format, such as logical rules or ontologies, for further reasoning and explanation.

An example of Neuro -> Symbolic integration could be in the domain of autonomous driving. A neural network-based perception system could first be trained to detect and classify various objects, such as vehicles, pedestrians, and traffic signs, from sensor data. The learned representations from this neural network could then be converted into a symbolic knowledge base, representing the detected objects, their properties, and the relationships between them. This symbolic knowledge could then be used for higher-level reasoning about the driving environment, planning safe maneuvers, and explaining the autonomous vehicle's decision-making process to human operators.

An example of using (Hassan, 2022) neuro-symbolic approach for automating medical report generation, highlighting the potential of combining neural networks and symbolic AI to achieve more reliable and interpretable medical decision support systems.

Which approach is more realistic in the near future?

The Neuro -> Symbolic approach seems more realistic in the near future, as it builds upon the recent advancements in deep learning and neural networks, which have demonstrated impressive performance on a wide range of tasks. By leveraging these neural capabilities and then translating the learned representations into symbolic forms, the Neuro -> Symbolic approach can potentially combine the strengths of both paradigms, enabling more interpretable and explainable AI systems.

The Symbolic Neurosymbolic approach, while promising, still faces challenges in seamlessly integrating symbolic and neural components, as well as in effectively learning the symbolic knowledge from data. Significant research is still needed to address issues like knowledge representation, reasoning, and the efficient interaction between the symbolic and neural components.

In the **near future**, we are likely to see more practical applications and deployments of **Neuro -> Symbolic systems** (), as the translation from neural to symbolic representations becomes more mature and the benefits of this integration become more apparent. For instance, IBM's work on neuro-symbolic AI focuses on developing novel architectures and techniques that combine the strengths of neural networks and symbolic AI, with applications in areas like computer vision, reasoning, and explainable AI. However, the Symbolic Neurosymbolic approach remains an important long-term goal, as it holds the potential to achieve more general and robust artificial intelligence by truly unifying the strengths of both symbolic and neural methods.

References:

Alcaraz, F., Fresno, V., Marchand, A. R., Kremer, E. J., Coutureau, E., & Wolff, M. (2018). Thalamocortical and corticothalamic pathways differentially contribute to goal-directed behaviors in the rat. *eLife*, 7, e32517. <https://doi.org/10.7554/eLife.32517>

Anticevic, A., Cole, M. W., Repovs, G., Murray, J. D., Brumbaugh, M. S., Savic, A. M. W. A., ... & Glahn, D. C. (2014). Characterizing thalamo-cortical disturbances in schizophrenia and bipolar illness. *Cerebral cortex*, 24(12), 3116-3130.

Goyal, A., & Bengio, Y. (2020). Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*.

Kipf, T. N., Fetaya, E., Wang, K. C., Welling, M., & Zemel, R. (2018). Neural relational inference for interacting systems. *arXiv preprint arXiv:1802.04687*.

Liu, J., Qi, Y., Meng, Z., & Fu, L. (2021). Self-learning Monte Carlo method. *Physical Review E*, 99(4), 043301.

IBM - Neurosymbolic AI, <https://research.ibm.com/topics/neuro-symbolic-ai>

Hassan, M., Guan, H., Melliou, A., Wang, Y., Sun, Q., Zeng, S., Liang, W., Zhang, Y., Zhang, Z., Hu, Q., Liu, Y., Shi, S., An, L., Ma, S., Gul, I., Rahee, M.A., You, Z., Zhang, C., Pandey, V.K., Han, Y., Zhang, Y., Xu, M., Huang, Q., Tan, J., Xing, Q., Qin, P., & Yu, D. (2022). Neuro-Symbolic Learning: Principles and Applications in Ophthalmology. *ArXiv*, abs/2208.00374.