AI & SOCIETY
https://doi.org/10.1007/s00146-024-02107-7

OPEN FORUM

Commonsense for AI: an interventional approach to explainability and personalization

Fariborz Farahmand[1]

Received: 25 April 2024 / Accepted: 14 October 2024

Abstract
AI systems are expected to impact the ways we communicate, learn, and interact with technology. However, there are still major concerns about their commonsense reasoning, and personalization. This article computationally explains causal (vs. statistical) inference, at different levels of abstraction, and provides three examples of how we can use *do*-operator, a mathematical operator for intervention, to address some of these concerns. The first example is from an educational module that I developed and implemented for undergraduate engineering students, as part of an educational research project with the US National Science Foundation. For the first time, to the best of my knowledge, 117 students could successfully use *do*-operator in a cybersecurity investment decision, according to Bloom's learning taxonomy. Gender did not make a significant difference in the students' performance, according to the Mann–Whitney U test. The second example explains using *do*-operator in assessing the effectiveness of intelligent tutoring systems, ITS, in receiving higher grades. The third example sheds light on combining online learning and offline learning, in reinforcement learning, to find the optimal policy that maximizes reward. To shed light on future research on explainability and personalization, I offer two recommendations: 1- Learn like System 2, the conscious learner (based on Bengio's proposal for deep learning 2.0), and 2- Preference, a process, not an object (based on preference analysis of 25,646 registrants, entities and individuals purchasing domain names). In conclusion, this article contributes to achieving the goal of human–AI: Machines that think that learn and that create.

Keywords Artificial intelligence · Commonsense reasoning · Intelligent systems · Learning

1 Introduction

"Despite its success, statistical learning provides a rather superficial description of reality that only holds when the experimental conditions are fixed" (Schölkopf et al. 2021). For example, statistical learning assumes identically and independently distributed data (iid), i.e. test data have the same distribution as the training data. But "real data arrives to us in a form which is not iid, and so in practice what many practitioners of data science or researchers do when they collect data is to shuffle it to make it iid. Nature does not shuffle data, and we should not" (Goyal and Bengio 2022). That is, we need to be able to formally represent "changes: consequence of an intervention on few causes or mechanisms" (Bengio 2020), e.g., changes of data distributions.

Causal reasoning, resting on interventions, enables us to formally represent changes. It allows us to decompose knowledge of the world into pieces and build abstract models of how the world works. It explains how changes in one variable, the cause, contributes to a change in the value of another variable, the effect, and generates commonsense explanations.

This is different from most of the existing machine learning and AI inference models that are based on statistical correlations (positive/negative/uncorrelated relations) with no causal implications. For example, an increase in ice cream sales could be correlated with an increase in violent crime. But this is not because ice cream causes crime. It is because both ice cream sales and violent crime are more common in hot weather (i.e., a confounder).

This article sheds light on the computational applications of causal (vs. statistical) reasoning in addressing various

✉ Fariborz Farahmand
fariborz@ece.gatech.edu

[1] School of Electrical and Computer Engineering, Georgia Institute of Technology, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, Georgia 30332-0765, USA

Published online: 09 November 2024

issues with AI systems. Following a discussion of related work, I present some of the results of an educational research project with the US National Science Foundation. I also present example applications of causal learning in addressing issues with intelligent tutoring systems and reinforcement learning. Then, I offer two recommendations for the future AI research on explainability and personalization, and present conclusions.

2 Related work

"If the AI behavior cannot be explained, then it will be difficult to trust its conclusions, and if the training inputs do not adequately represent the real environment, then we cannot have confidence that it works correctly across all inputs that may occur in practical use" (Laplante and Kuhn 2022). But despite receiving increasing attention, and an in-depth work on explainable AI, e.g., (Chou et al. 2022; Buijsman 2022), the existing AI systems still struggle with tasks that involve higher levels of reasoning. For example, when asked "what considerations are involved when transporting Egypt across the Golden Gate Bridge? ChatGPT failed to recognize that Egypt is a country and could not be transported across the Golden Gate Bridge, and produced a paragraph about weight, width, speed, and environment" (Denning 2023).

Explaining concepts and observed phenomena in the form of causal mental models is as old as human cognitive systems. But to automate commonsense explanation, e.g., in ChatGPT, we need to express causal inference with algorithms or formal rules that take representations as input and produce representations as output. Pearl (2009) made a fundamental contribution to causal inference by developing *do*-calculus, a calculus for probabilistic and causal reasoning, and *do*-operator, the operator that allows to intervene–interventions of the form $do(X = x)$, which forces the variable X to take *only* the value x, and have no other immediate effect. The *do*-operator is specifically helpful when we cannot conduct randomized controlled trials, RCT. For example, in the context of education, when conducting an RCT requires changing classroom sizes and substantial resources. Pearl's causal reasoning has three levels, as summarized in Table 1, and briefly explained below.

- Level 1 (association). It invokes purely statistical relationships, defined by the naked data. For example, it uses $\varepsilon P(y|x) = p$" that stands for: "The probability of event $Y = y$, given that we observed event $X = x$ is equal to p." Queries at this layer are placed at the bottom level in the hierarchy because they only present associations and not causal relations.
- Level 2 (intervention). This level ranks higher than association because it involves not just observing "what is" but changing what we observe. For example, it uses "$P(y|do(x), z)\varepsilon$ that stands for: "The probability of event $Y = y$, given that we intervene and set the value of X to x and subsequently observe event $Z = z$."
- Level 3 (counterfactuals). This is the highest level of the hierarchy because it subsumes interventional and associational questions. For example, it uses "$P\left(y_x|x\prime, y\prime\right)\varepsilon$ that stands for: "The probability that event $Y = y$ would be observed had X been x, given that we actually observed X to be $x\prime$ and Y to be $y\prime$." (Pearl 2019)

I argue that such three-level hierarchy is beneficial to AI research and education, as it corresponds to Turing proposal to classify a cognitive system in terms of queries it can answer. Specifically, *do*-operator, using graph theory, can help explain crosscutting concepts in different computational forms—languages familiar to the engineers and computer scientists. This can significantly help engineering and computer science research and education with answering questions and reasoning at different levels of abstractions, using formal semantics and graphical representations.

Current methods of personalized decision-making are mainly based on *average* treatment effect, ATE, and propensity score of a population. This could be problematic because personalized models and individual (personalized) treatments are expected to target the behavior of an individual, not the population that may not necessarily resemble that individual. That is, the propensity score of a population may not be the same as an individual's propensity.

Table 1 Pearl's causal hierarchy

Level	Typical Activity	Typical Questions
1. Association: $P(y\|x)$	Seeing (observing a certainphenomenon unfold)	What is? How would seeing Xchange my belief in Y ?
2. Intervention: $P(y\|do(x), z)$	Doing (acting in the world to bring about some state of affairs)	What if? What if I do X?
3. Counterfactuals: $P\left(y_x\|x\prime, y\prime\right)$	Imagining (thinking about alternative ways the world could be)	Why? Was it X that caused Y? What if I had acted differently?

The reality is that, to develop personalized decision-making models, we need interventional expressions for an individual, and this cannot be done by ATE and passive observations from a dataset alone, regardless of how big the dataset is. Formally speaking, the average treatment is expressed as:$ATE = E(Y|X = 1) - E(Y|X = 0)$. We can advance this, using *do*-operator, and express treatment effectiveness by the Individual Treatment Effect as: $ITE = E(Y|do(X = 1)) - E(Y|do(X = 0))$. An advantage of *ITE* over *ATE* is that in*ATE*, we need to have RCT, and the assumption of independence of other factors is implicit. But in *ITE*, we do not necessarily need to have RCT, and we explicitly denote, using *do*-operator, that treatment status is independent of other factors.

We can extend this approach by including counterfactual reasoning and bounds (point of estimate) on individual-level causation to assess the probability of causation for personalized decision-making. The counterfactual view of causation, y would not have occurred if it was not for x, can help with the assessment of probability of causation. This is like using the legal concept of *but for* in settings such as the plaintiff must prove that y would not have occurred but for x. Pearl adapted this view to define the probability of necessity, *PN*, as the: "probability that event y would not have occurred in the absence of event x, given that x and y did in fact occur" (Pearl 2009). He developed conditions for which *PN* can be learned from data, and how data from both experimental and nonexperimental studies can be combined to yield information that neither study alone can provide. He also developed bounds on individual-level causation. For example, assuming identifiability (situations where interventional distributions can be obtained from the given data), we can define a lower bound for *PN* as:

$$PN \geq \left\{ \max 0, \frac{P(y) - P(y|do(x\prime))}{P(x, y)} \right\}$$

This following section briefly explains my experiments where I used PN and the lower bound in a real-world cybersecurity scenario.

3 Experiments

Here, I present some of the results of an educational research project with the US National Science Foundation, where I created and implemented curriculum modules that are focused on AI in cybersecurity and are infused with real-world scenarios (Farahmand 2021). Following receiving an institutional review board, IRB, approval to allow data gathering from volunteer students, 117 engineering undergraduate students, with no related background, voluntarily completed their homework, following a 40 min lecture. Our first module consisted of three parts.

Part 1 of the module included a quick review of the basic statistical and probabilistic concepts that students needed to understand the rest of the module. It also included examples on how standard Bayesian inference can be used in the assessment of suspects in a cybercrime investigation.

Part 2 of the module introduced intervening (vs. conditioning) and causal reasoning, i.e., reasoning for situations where one intervenes in the world, thereby interfering in the natural course of the events. Key to this part was the fundamental distinction between regression coefficients and structural parameters, and how students can use both to predict causal effects in linear models, and work with Pearl's *do*-calculus, a general calculus for identifying causal effects. For example, use $do(X = x)$ to force the variable X to take the value x, having no other immediate effect. Part 2 explained that a causal model can be interpreted as a Bayesian network, which in addition to answering probability queries, can also answer intervention queries; and that the answer to an intervention query $P(Y|do(z), X = x)$ is not generally the same as its corresponding probability query $P(Y|Z = z, X = x)$.

Part 3 of the module introduced the concept of counterfactuals–what would have happened had we chosen differently at a point in the past. Discussions followed on how to compute counterfactuals, estimate their probabilities (e.g., probability of necessity that captures the legal criterion of "but for"), and how to use counterfactuals to answer practical questions in cybersecurity (e.g., cyber attribution).

In part 3, I also introduced the three rules of *do*-Calculus rules. First, in simple terms, I explained *do*-calculus, a calculus for probabilistic and causal reasoning (in Pearl's words, "machinery of causal calculus") is an axiomatic system for replacing probability formulas containing the *do*-operator with ordinary conditional probabilities that uses three rules:

- Rule 1 helps us to ignore observations. It says when we observe a variable W that is irrelevant to Y (possibly conditional on other variables Z), then the probability distribution of Y will not change.
- Rule 2 helps us to exchange actions with observations. It says if a set Z of variables blocks all backdoors from X to Y, i.e., any path from X to Y that starts with an arrow pointing into X, then conditional on Z, $do(X)$ is equivalent to observe (X).
- Rule 3 helps us to ignore actions. It says we can remove $do(X)$ from $P(Y|do(X))$ in any case when there are no causal paths from X to Y. That is, if we do something that does not affect Y, then the probability distribution of Y will not change.

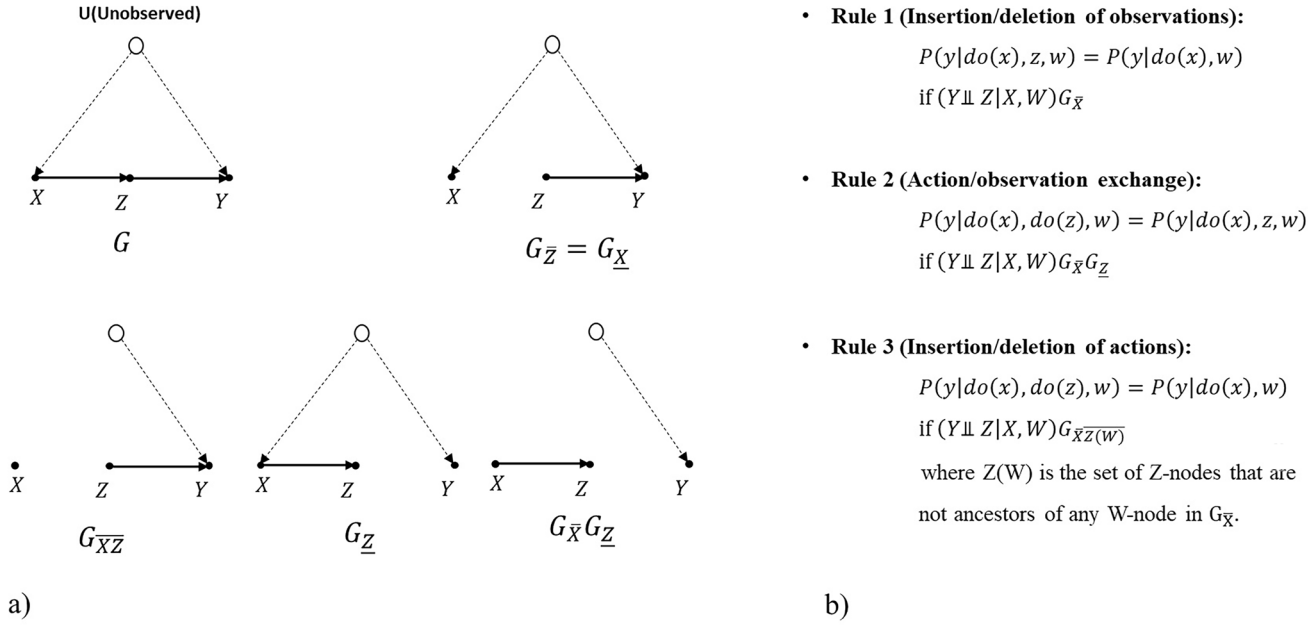

Fig. 1 **a** Subgraphs of G used in the derivation of causal effects, **b** Rules of *do*- calculus

Then I provided a graphical explanation (see Fig. 1 next page) of *do*-calculus rules to show how to apply the rules of *do*-calculus and *do*-operator to untangle causation in a cybersecurity investment decision, to answer a "what if" question (level 2 of causal hierarchy, see Table 1).

In Fig. 1, X, Y, Z and W are arbitrary disjoint sets of nodes in a causal directed acyclic graph$G$, as depicted in Fig. 1a. Here, an arrow from one variable to another indicates that the first variable causes the second—that is, the value of the first variable is part of the function that determines the value of the second. Therefore, the second variable depends on the first for its value. $G_{\overline{X}}$ denotes the graph obtained by deleting from G all arrows pointing to nodes in$X$, and $G_{\underline{X}}$ denotes the graph obtained by deleting from G all arrows emerging from nodes in$X$.$G_{\overline{X}\underline{Z}}$, represents the deletion of both incoming and outgoing arrows. Figure 1b explains the three rules of *do*-calculus to help with eliminating the *do*-operators from the query expression and working with the observational data. For example, Rule 3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of$Y = y$.

All three parts of the first module included computational examples of the applications of causal inference in either tangible, real-life situations, or in real-world cybersecurity situations. Following completing part 3, students voluntarily completed the following example homework.

4 Example homework problem: applying *to do*-operator to a cybersecurity investment decision

In this problem, students learned how to use *PN* and the lower bound in a real-world cybersecurity scenario. Assume board members of a company, Board, need to choose between two treatments to protect the company from certain malicious attacks. Treatment 1, offered by the chief security officer, CSO, who recommends firewall plus antivirus software, and Treatment 2, offered by the chief financial officer, CFO, who argues that purchasing antivirus software is unnecessary and additional cost, and recommends a different treatment: The firewall alone.

The following is a summary of the experimental and observational data that is available to the Board:

- X: Treatment ($x\prime$ representing CFO's treatment, i.e., firewall alone, x representing CSO's treatment: i.e., firewall plus antivirus),
- Y: Protecting the company ($y\prime$ representing company unprotected,y representing company protected),
- $P(y\prime) = 0.3$,
- $P(x\prime|y\prime) = 0.7$,
- $P(y|do(x)) = 0.39$, and
- $P(y|do(x\prime)) = 0.14$.

The Board needs to determine if the CSO's treatment is likely necessary to protect the company from certain malicious attacks. That is, Board needs to see if *PN* is more probable than not, using the lower bounds on *PN*,

and assess if $PN \geq \frac{P(y)-P(y|do(x\prime))}{P(x,y)}$ holds. Using standard probability axioms, the right-hand side of the inequality can be written as:

$$= \frac{P(y) - P(y|do(x\prime))}{P(y|x)P(x)}$$

$$= \frac{P(y) - P(y|do(x\prime))}{\left(1 - \frac{P(x|y\prime)P(y\prime)}{p(x)}\right)P(x)}$$

$$= \frac{P(y) - P(y|do(x\prime)}{P(x) - P(x|y\prime)P(y\prime)}$$

Using the available data, the Board finds the following:

$$PN \geq \frac{0.7 - 0.14}{1 - (1 - 0.7) \times 0.3}, or = 0.62 > 0.5$$

Since *PN* is greater than 0.5, the Board can conclude that the CSO's recommendation is necessary for company's protection from certain malicious attacks.

To assess students' learning, using *do*-operator, I used Bloom taxonomy (Anderson et al. 2001), as in computer science education, like many other disciplines, Bloom taxonomy is widely used for students' learning assessment (Fuller et all 2007; Gluga et al. 2012; Moore et al. 2023; Ng et al. 2021; Sulmont 2019). The learning assessment of the Joint Task Force on Computing Curricula, a collaborative effort by the Association for Computing Machinery (ACM), IEEE-Computer Society (IEEE-CS), and Association for Advancement of Artificial Intelligence (AAAI) is also aligned with the Bloom taxonomy (Kumar et al. 2023).

The Bloom's six levels of learning that I used in our assessment were 1-Remember, 2-Understand, 3-Apply, 4-Analyze, 5-Evaluate, and 6-Create. All participants reached level 4. That is, they were able to remember, understand, apply, and analyze the lecture materials in answering the homework questions. Fifty-three percent of participants reached levels 5 and 6. That is, in answering their homework questions, they were able to work with the *do*-operator to evaluate and justify a decision and put elements together in a creative new way. The average and the highest score for the male students were 85 and 100, and for the female students were 81 and 97, respectively. Gender did not make a significant difference in the students' performance, according to the Mann–Whitney U Test; p-value was found as 0.25, and the result was not significant at $p < 0.05$. These results are significant, as the 117 students who participated in this study and voluntarily completed their homework, using *do*-operator, had no related background.

5 Example applications

The following two examples shed light on the applications of *do*-operator and causal learning in addressing issues with intelligent tutoring systems and reinforcement learning.

5.1 Applying *do*-operator to assess the effectiveness of intelligent tutoring systems

This example sheds light on using *do*-operator and the three levels of causal reasoning in assessing the effectiveness of intelligent tutoring systems, ITS, in receiving higher grades.

Assuming to succeed in a new program, Teacher 1, who strongly believes in the power of encouragement, has encouraged students to work harder by doing more homework, using an intelligent tutoring system, ITS, that provides worked examples. However, Teacher 2 presents a counterargument: Program's success is substantive, achieved mainly due to the unique features of the curriculum covered, and the increase in homework efforts cannot alone account for the success observed. But Teacher 2 does not provide any data to support his counterargument. To respond, Teacher 1, using graph theory and *do*-operator, can explain his argument at three levels, as shown in Fig. 2.

At level 1, Teacher 1, using conditional probability, assesses the degree of association between students receiving higher grades, and worked examples, using ITS. At this level, Teacher 1 explains his argument as: The probability of receiving higher grade, given the observation of students using ITS. This is the common method of reasoning, used by machine learning research, that invokes purely statistical relationships, defined by the naked data.

At level 2, Teacher 1, using *do*-operator, does intervention (vs. conditioning) to simulate an RCT. Intervention ranks higher than association because it involves not just seeing "what is" but changing what we see. At this level, Teacher 1 explains his argument as: The probability of receiving higher grade, given the students were made to use ITS.

At level 3, Teacher 1 uses counterfactual reasoning. This is the highest level of causal reasoning because it subsumes interventional and associational questions. At this level, Teacher 1 explains his argument as: The probability of not receiving higher grades, given the students received higher grades and did not use ITS.

DoWhy (2023), a Python library, and DAGitty (Textor 2023), an R package, are examples of the tools that are publicly available and used by researchers and practitioners with interest in do-operator.

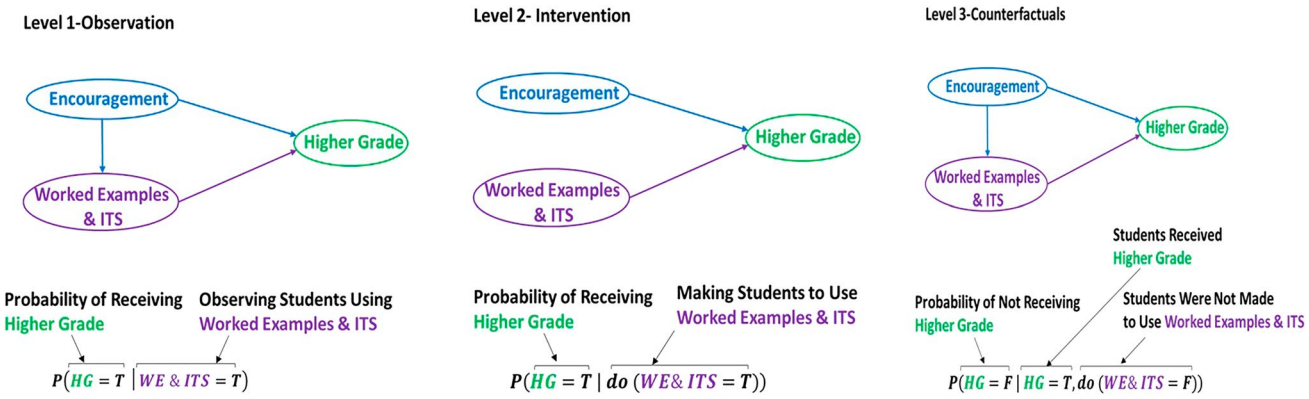

Fig. 2 Three levels of causal reasoning, using conditional probability (level 1) and *do*- operator (levels 2 and 3)

5.2 Applying *do*-operator to combine online and offline learning

This example sheds light on combining online learning and offline learning, in reinforcement learning, RL, to find the optimal policy that maximizes reward.

RL methods that learn behaviors based on feedback from the environment are closer to causality research than the machine learning mainstream. RL methods can be categorized into two groups: 1) Online learning, where agents perform experiments themselves, and 2) Offline or off-policy learning, where agents learn from other agents' actions. Offline learning is specifically helpful when conducting RCT is not feasible.

RL "uses the formal framework of Markov decision processes to define the interaction between a learning agent and its environment in terms of states, actions, and rewards. This framework is intended to be a simple way of representing essential features of the artificial intelligence problem. These features include a sense of cause and effect, a sense of uncertainty and nondeterminism, and the existence of explicit goals." (Sutton and Barto 2018) For example, in assessing ITS performance, we can consider problems in the framework of Markov decision process, where an agent collects rewards over time by performing actions in an environment, as briefly described below.

Assume we plan to choose the optimal policy for an ITS. We can define a Markov decision process as follows:

- **States:** States that can be defined based on student's performance, and level of knowledge of the subject
- **Actions:** Actions that the agent can take to change the student's state, e.g., presenting worked examples
- **Rewards:** Rewards that student receive based on performance, e.g., "Excellent", "+1"
- **Training:** Observing actions
- **Goal:** Finding the optimal policy that maximizes reward

However, there are some challenges with interpretability of reinforcement models. For example, offline learning methods "have a long history of using importance sampling and yet still are not well understood". (Sutton and Barto 2018).

I argue that like deep learning methods that have benefited from large datasets and methods that could scale to large amounts of data, RL methods can benefit from *do*-operator and graph theory. Such a combination enables us to draw conclusions about new policies, by combining observations and knowledge about the data-generating mechanisms. Such combination also enables us to benefit from both online and offline learning in a formal setting and systematically combine the results of our limited interventional studies with our diverse prior experiences, and observational data (logged data).

Consider our goal of finding the optimal policy that maximizes reward. It can formally be expressed as: Learning a policy π s.t. sequence of actions $\pi(.) = \left(X_1, X_2, \ldots, X_n\right)$ maximizes reward$E\pi(Y|do(X))$. To achieve this goal, we can do a combination of online learning by an agent, who performs experiments itself with input: experiments $\left\{\left(do\left(X_i\right), Y_i\right)\right\}$ and learn$P(Y|do(X))$, and offline learning by an agent, who learns from other agents' actions with input: samples $\left\{\left(do\left(X_i\right), Y_i\right)\right\}$ and learn$P(Y|do(X))$. Such combinations allow agents to systematically combine the observations and interventions it's already collecting to construct an equivalence class of causal models (Zhang and Bareinboim 2020). Such combinations are also the roots of the concept of transfer learning in reinforcement learning – the use of observational (or offline) data to aid in the performance of an agent in an experimental (or online) setting.

One of the first works on using causality to combine observational and experimental data came in Forney et al. 2017, which proposed a counterfactual approach to the fusion of the two types of data within a multi-armed bandit setting. In this work, the authors approached the problem of

unobserved confounders when relating the two data sources by using a counterfactual quantity called the effect of the treatment on the treated (ETT) to estimate the causal effect of the action upon the reward, which they combined with empirically observed results from experimentation. Moving beyond the setting of multi-armed bandits, Gasse et al. 2021 followed a similar approach to Zhang and Bareinboim 2020, by creating causal bounds from a combination of observational and interventional data. Wang et al. 2021 introduced a value iteration approach to improve sample efficiency by combining online and offline data in settings with confounded observational data. They proposed a method called de-confounded optimistic value iteration (DOVI), which uses causal inference to adjust for confounded behavior observed in offline data. Like Zhang and Bareinboim 2017, Kamath et al. 2023 showed that informative observational data provably improves efficiency in the online setting.

6 Recommendations for future research

Here, I offer two recommendations for future research on explainability and preference analysis in AI systems.

6.1 Learn like system 2, the conscious learner.

One aspect of preference in AI research has been studied under (inductive) biases, factors that lead a learner to favor one hypothesis over another that are independent of the observed data. "A remaining important question for AI research aiming at human-level performance then is to identify inductive biases that are most relevant to the human perspective on the world around us" (Goyal and Bengio 2022).

To address this need, Bengio based on Kahneman's System 1 and 2 thinking (Kahneman 2011), has proposed to move from current deep learning, DL, to DL 2.0, where System 1 is unconscious, fast, intuitive, and emotional, and System 2 is conscious, slower, more deliberative, and more logical. Bengio argues that the origin of explainability issues in machine learning is in taking a System 1 approach to learn from data.

Figure 3a shows a situation where we are driving in a familiar neighborhood, e.g., from work to home, where we don't need any map, and don't need to think about turning to left or to right. It is off the top of our head! This is System 1—how the current machine learning answers queries. Here, there is knowledge about how to solve problems, but it is implicit. That is why we have explainability issues in machine learning.

In contrast, Fig. 3b shows driving in an unfamiliar neighborhood, where we need to become slower, perhaps consult with Google Map, and need to deal with what could go wrong. With System 2, we are not only conscious, but also are able to express our thinking verbally, something that we are unable to do with current machine learning. As illustrated in Fig. 3b, you ask the other person not to talk to you because you must focus your thoughts on your driving. This conscious processing is slower, like you need to think carefully before you act. It is logical. You can explain to people why you are making such choices. This is like a situation when we are designing an algorithm. Our mind processes information sequentially and the knowledge that we manipulate is explicit. Therefore, we can explain to others why we did, or did not, do something. This is the kind of capability that machine learning needs to have, so it can help AI systems to manipulate the semantic concepts that we may even know already.

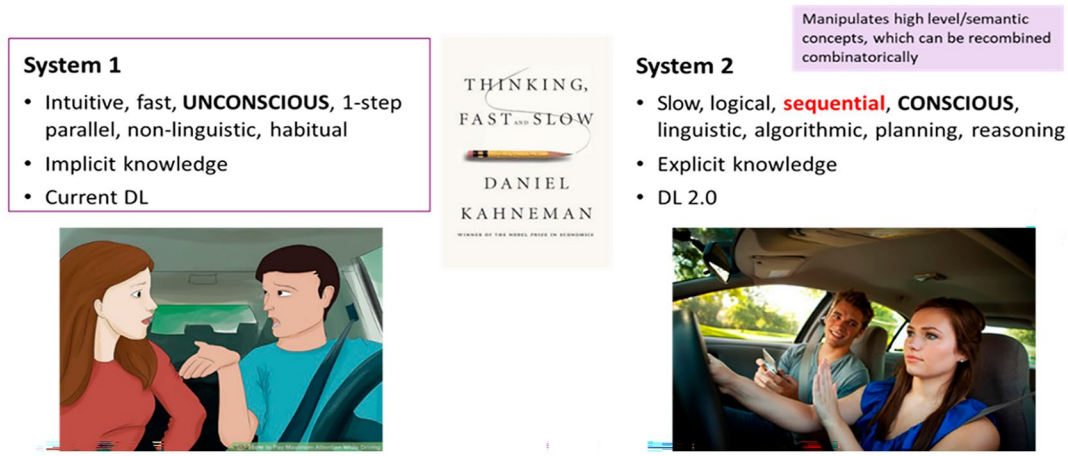

Fig. 3 **a** System 1 versus **b** System 2 thinking (Source: Bengio 2020, used with permission)

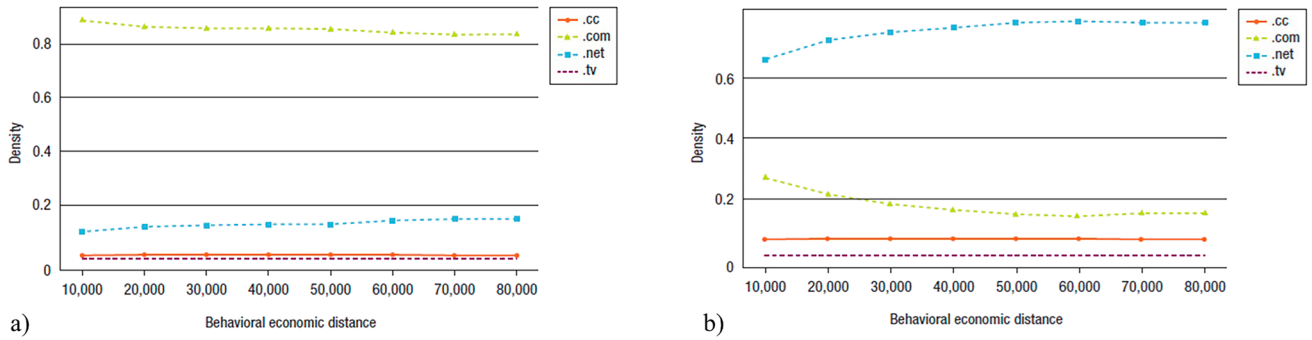

Fig. 4 Depicting density graphs that give the percentage of registrants choosing **a** TLD as a function of behavioral economic distance when [keyword].com is a) available, and **b** unavailable

The kind of explainability that Bengio and the AI community desire for DL 2.0 needs relaxing assumptions such as the identically and independently distributed, iid, data. It requires the learner to discover models that capture the effect of interventions and distribution changes that can be realized by performing level 2 and 3 actions, in Pearl's hierarchy, using *do*-operator.

6.2 Preference, a process, not an object.

Human preferences are not objects that can be identified. They are often constructed in the process of elicitation, and the AI community needs to investigate the constructive nature of human preferences–opposed to stable preferences. As Kahneman (2011) explained: "preferences are frame-bound rather than reality-bound", and "preferences between the same objective outcomes reverse with different formulations". Kahneman's argument sharply contradicts the axioms of expected utility theory, the dominant decision theory in AI and machine learning.

I studied Kahneman's argument in the context of domain name registration where.com is always assumed to be the most preferred top-level domain, TLD, name (Farahmand 2017). I studied how online users choose domain names, looking at decisions from 25,646 registrants (the entities and individuals purchasing domain names) from a set that included the.com,.net,.cc, and.tv TLDs. I developed a behavioral-economic distance metric, using behavioral economics and decision science research, and evaluated how the similarity effect—a type of contextual effect—influences domain name choice. My results indicate that preference depends not only on being or not being a.com, but also on what choices the registrant was given and the presentation context. This goes against independence of irrelevant alternatives (IIA) principle of expected utility theory: people have a stable, well defined, and discernible order of preferences and always choose the course of action that maximizes their preferences (utilities). That is, in contrast to expected utility theory, we cannot assume that people have a stable, well defined, and discernible order of preferences and always choose the course of action that maximizes their preferences.

The registration records contained three main parts: 1) Keyword, the original query by the registrant that specified the domain name without a TLD extension, 2) Registered domain, the domain name and the TLD selected for registration, and 3) Similarity score, a measure of the suggestion's similarity to the keyword. Figure 4 shows registrant preference as a function of behavioral-economic distance metric. Figure 4a depicts density when [keyword].com is available; Fig. 4b depicts density when it is not, showing the move from.com to.net as the preferred TLD when keyword.com. became unavailable.

7 Conclusions

This article sheds light on how we can advance personalization in AI systems to causal (vs. purely statistical) learning and presents computational examples and recommendations to illustrate how causal learning can lead to a human-like explainability. It formulates how to use casual reasoning, at different levels of abstraction, to address AI research issues, e.g., use commonsense explanation in conversational AI. It proposes to advance RL research by systematically combing observations and interventions, using *do*-operator, and goes beyond ATE, by bounding the entire distribution of individual causal effects.

This article contributes to engineering and computer science education by advancing design thinking—formulating solution-based and user-centric rather than problem-based approaches. It presents measurable outcomes of introducing, for the first time to the best of our knowledge, a Turing-Award winning research (Pearl's *do*-calculus) to engineering education. This article also contributes to research on AI systems by integrating causal reasoning into personalization,

commonsense explanation, and achieving the goal of human–AI: Machines that think that learn and that create.

Author contributions The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Funding This material is based upon work supported by the National Science Foundation under award number 2041788.

Data availability The author confirms that all data generated or analyzed during this study are included in this article.

Declarations

Conflict of interest The author declares no conflict of interests.

References

Anderson L et al (2001) A Taxonomy for Learning, Teaching, and Assessing, A Revision of Bloom's Taxonomy of Educational Objectives. Longman, White Plains

Bengio Y (2020) Deep Learning for System 2 Processing, Presentation at the AAAI-20 Turing Award winners 2018, Feb. 9

Buijsman S (2022) Defining explanation and explanatory depth in XAI. Mind Mach 32:563–584

Denning P (2023) The Profession of IT Can Generative AI Bots Be Trusted? Commun ACM 66(6):24–27

DoWhy (2023) https://www.pywhy.org/dowhy/v0.10/. Accessed 10 July 2024

Farahmand F (2017) The importance of human information processing: a behavioral economics model for predicting domain name choice. Computer 50(9):67–74

Farahmand F (2021) Integrating Cybersecurity and Artificial Intelligence Research in Engineering and Computer Science Education. IEEE Secur Priv 19(6):104–110

Forney A, Pearl J, Bareinboim, E (2017) Counterfactual Data-Fusion for Online Reinforcement Learners. Proceedings of the 34th International Conference on Machine Learning, edited by Doina Precup and Yee Whye Teh, 70:1156–64. Proceedings of Machine Learning Research. PMLR.

Fuller U et al (2007) Developing a computer science specific learning taxonomy. ACM SIGCSE Bull 39(4):152–170

Gluga R, Kay J, Lister R, Kleitman S, Lever T (2012) Coming to terms with Bloom: an online tutorial for teachers of programming fundamentals. In Proceedings of the Fourteenth Australasian Computing Education Conference. 123-147-156.

Goyal A, Bengio Y (2022) Inductive biases for deep learning of higher-level cognition". Proceed Royal Soc. https://doi.org/10.1098/rspa.2021.0068

Kahneman D (2011) Thinking, Fast and Slow. Farrar Straus Giroux

Kamath U, Graham K, Naylor M (2023) Applied Causal Inference

Kumar AN et al (2023) Computer Science Curricula 2023. ACM Press, IEEE Computer Society Press and AAAI Press

Laplante P, Kuhn R (2022) AI Assurance for the Public – Trust but Verify, Continuously, IEEE 29th Annual Software Technology Conference (STC), 174–180

Moore S, Fang E, Nguyen H A, Stamper J (2023) Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom's Taxonomy. In Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23). Association for Computing Machinery, New York, USA, pp. 25–34

Ng DTK et al (2021) Conceptualizing AI literacy: An exploratory review, Computers and Education: Artificial Intelligence. Artificial Intell, Comput Educat. https://doi.org/10.1016/j.caeai.2021.100041

Pearl J (2009) Causality. Cambridge University Press

Pearl J (2019) The seven tools of causal inference, with reflections on machine learning. CACM 62(3):54–60

Schölkopf B et al (2021) Toward causal representation learning,". Proc IEEE 109(5):612–634

Sulmont E, Patitsas E, Cooperstock JR (2019) What is hard about teaching machine learning to non-majors? insights from classifying instructors' learning goals. ACM Trans Comput Educ. https://doi.org/10.1145/3336124

Sutton RS, Barto AG (2018) Reinforcement Learning. Second edition, The MIT Press, An Introduction

Textor J (2023) https://www.dagitty.net/. Accessed 10 July 2024

Wang L, Zhuoran Y, Wang Z (2021) Provably Efficient Causal Reinforcement Learning with Confounded Observational Data". In Advances in Neural Information Processing Systems, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and. J. Wortman Vaughan. 34:21164–21175

Y-L. Chou Y-L, et al (2022) Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion 81:59–83

Zhang J, Bareinboim E (2017) Transfer Learning in Multi-Armed Bandits: A Causal Approach. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI.17:1340–46

Zhang J, Bareinboim E (2020) Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach. Proceedings of the 7th International Conference on Machine Learning. 119:11012–11022