Formalizing Cognitive Biases in Medical Diagnostic Reasoning

Dominik Battefeld*, Stefan Kopp

Social Cognitive Systems Group, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany

Abstract

This paper presents preliminary work on the formalization of three prominent cognitive biases in the diagnostic reasoning process over epileptic seizures, psychogenic seizures and syncopes. Diagnostic reasoning is understood as iterative exploration of medical evidence. This exploration is represented as a partially observable Markov decision process where the state (i.e., the correct diagnosis) is uncertain. Observation likelihoods and belief updates are computed using a Bayesian network which defines the interrelation between medical risk factors, diagnoses and potential findings. The decision problem is solved via partially observable upper confidence bounds for trees in Monte-Carlo planning. We compute a biased diagnostic exploration policy by altering the generated state transition, observation and reward during look ahead simulations. The resulting diagnostic policies reproduce reasoning errors which have only been described informally in the medical literature. We plan to use this formal representation in the future to inversely detect and classify biased reasoning in actual diagnostic trajectories obtained from physicians.

Keywords

Diagnostic reasoning, Cognitive bias, Cognitive model, POMDP, Bayesian network, Epilepsy, CDSS.

1. Introduction

Medical diagnostic reasoning is exposed to incomplete and uncertain information. No physician will ever be able to know *everything* about a patient. Likewise, no physician can be 100% certain that the chosen diagnosis - given the limited amount of information - is correct. Medical experts face the problem of hypothesis weighting nearly every day and problem complexity scales with the quality of evidence. While modern laboratory tests may deliver robust, isolated hints towards one diagnostic option or the other in some medical domains, diagnosing transient loss of consciousness remains a challenging and highly prevalent problem. "*Almost 10% of people will experience at least one seizure over a lifetime*" [1, p. 1]. These seizures can originate from various causes ranging from one-time explanations like sleep deprivation and mental overload to potentially life-long diagnoses like epilepsy. The three main differential diagnoses to consider are an epileptic seizure, a syncope and a psychogenic non-epileptic seizure (PNES) [2, p. 96] and to prioritize one cause over the other can be seriously complex. At the time of writing, no

8th Workshop on Formal and Cognitive Reasoning, September 19, 2022, Trier, Germany

*Corresponding author.

dbattefeld@techfak.uni-bielefeld.de (D. Battefeld); skopp@techfak.uni-bielefeld.de (S. Kopp)

https://techfak.uni-bielefeld.de/~dbattefeld/ (D. Battefeld); https://techfak.uni-bielefeld.de/~skopp (S. Kopp)

0000-0002-5480-0594 (D. Battefeld); 0000-0002-4047-9277 (S. Kopp)

gold standard tests exists to confidently confirm or rule out epilepsy as a diagnosis candidate [3, p. 144]. Additionally, the isolated presence or absence of a semiological feature cannot warrant a diagnostic decision towards or against epilepsy [3]. Information is mainly obtained through subjective, personal dialogue rather than objective test results [4] and consequently, evidence quality suffers from inaccurate memory retrieval and miscommunication. The diagnosis of seizure-like events is thus a prime example of the initially mentioned incompleteness and uncertainty inherent to medical decisions. Today, the most successful approach to diagnose epilepsy and thus solve the medical decision problem is efficient knowledge exploration, i.e., detailed history taking, critical eyewitness report analysis and conservative test result interpretation [3, 5, 4]. By that, experts learn to maneuver in the space of retrievable medical evidence. The sequence of questions they ask, features they query and tests they issue defines one possible diagnostic trajectory through this space and the overall goal is not to cover everything but everything *relevant*.

In the endeavour to ease diagnostic pressure on physicians in highly complex domains like these, clinical decision support systems (CDSS) provide assistance in various ways for both doctors and patients [6, 7]. Some try to increase the certainty of information by tracking and storing features about specific seizures [8, p. 98]. Others facilitate information integration. They act as differential diagnosis generators which accept a set of present medical findings and test results and output candidate diagnoses along with their respective probability [9, 10, 11]. What these state-of-the-art systems lack is an explicit model of their user. They can be queried for additional information but are agnostic to the reasoning process in the physician's mind. Thus, some argue that the development paradigm of clinical decision support systems should shift from resembling an additional uncertain source of information (i.e., "Is the system output correct?") towards facilitating the process itself in a clinical reasoning support system [12]. This is especially evident in the fact that most diagnostic errors are not rooted in insufficient medical knowledge or expertise, but rather in structural causes like time pressure and cognitive biases like premature closure [13, 14]. Additionally, reasoning errors are especially prevalent in areas of high subjectivity like epilepsy. It suffers from an interobserver variation, where multiple practitioners don't agree in their diagnosis of the same patient. Misdiagnosis rates are estimated around 23% or even higher in everyday practice [5].

Taken all together, the goal of this paper is to formally define the relation between a specific cognitive bias (i.e., premature closure, confirmation bias and availability bias) and the erroneous knowledge exploration trajectory leading up to the diagnostic reasoning error that may result from it. Premature closure describes the tendency to submit an unjustified diagnosis too early while a confirmation bias leads to a skewed interpretation of observations [15]. Even conflicting evidence is seen as in line with the current beliefs. An availability bias causes physicians to favor what is familiar by overestimating the likelihood of hypotheses that "*readily come to mind*" [15, p. 777]. We capture this faulty reasoning in a partially observable Markov decision process (POMDP) by altering rewards and observation likelihoods during belief update and policy computation. Here, the idea of a biased policy computation is based on the premise that planning always involves a predictive component of what happens next after executing some action. If the understanding of the world is in itself biased, mental action execution during planning will lead to biased world states which will manifest in biased action assessment and thus in a biased policy. To explain this approach in more detail, Section 2 presents related

work on the formalization of diagnostic processes and cognitive biases in diagnostic reasoning. Section 3 introduces the cognitive modeling work and how each bias can be integrated into it. Section 4 presents the resulting policies under each bias and Section 5 concludes with a discussion of the results, potential shortcomings and next steps.

2. Related work

In general, POMDPs "*model sequential decision making problems where the agent must act under partial observability of the environment state. POMDPs consider both uncertainty in action effect (i.e., transitions) and observations, which are usually incomplete and noisy information related to the state*" [16, p. 2]. They are are a well known and frequently used formalism to operationalize medical decision-making in various domains. In practice, Zhang et al. [17] model the diagnostic process and subsequent management of chronic diseases in the example domain of coronary heart disease by maximizing the patient's total expected quality-adjusted life years (QALYs). They focus on deriving structural properties like diagnostic thresholds and the optimal age for a screening. Similarly, Zhang et al. [18] propose a non-stationary POMDP to model medical screening for prostate biopsy referral decisions. Their model maximizes expected quality-adjusted life years as well and incorporates prostate-specific antigen test results as observations within the formalism. Li et al. [19] motivate the usage of a partially observable Markov chain to assess different screening strategies during colonoscopy. By optimizing parameters like initial screening age or screening frequency, they aim to increase the diagnostic accuracy of colorectal cancer. Arruda et al. [20] try to mathematically define a stochastic shortest path policy of tests to confirm or discard a disease hypothesis. Their approach uses Bayesian inference to turn an a priori disease probability into posterior probabilities that warrant either immediate treatment or a not-ill diagnosis. Partially observable Markov decision processes are also used to infer a suitable medical treatment policy. Here, Bazrafshan et al. [21] use a finite-horizon MDP to formalize the problem of planning chemotherapy and optimal drug administration to treat gastric and gastroesophageal cancers. Hauskrecht et al. [22] employ the POMDP framework "*to model and solve the problem of the management of patients with ischemic heart disease*" [22, p. 221]. Ibrahim [23] use the decision framework in two separate phases. A first POMDP model formalizes the physicians explorative process to understand how sensitive the patient reacts to a medication with warfarin. Then, a second MDP model is used to calculate the optimal treatment policy of the patient with warfarin based on the beliefs and sensitivities derived from the first stage.

We adopt these previous approaches in which the potential state of the patient forms the state space [17, 18, 19]. Sequential action selection is based on the current subjective belief in each possible patient state and obtained medical information updates the belief after each action using the Bayes calculus. Using a POMDP enables us to represent the current belief and its update in an efficient and transparent manner at each time step. We share the view of Arruda et al. [20] to frame a diagnosis as a stochastic shortest path problem of information retrieval actions up to a diagnosis submission as the point at which "sufficient justification" is reached. This connects to our idea of economically exploring the space of available medical knowledge to only query features that are relevant at this point in time and neglect redundant

or insignificant ones. This trade-off between acquiring more information to increase certainty and making a diagnosis as fast as possible to decrease suffering is implicitly represented in the policy computation.

Diagnostic errors induced by cognitive biases are an increasingly explored field of research. The shortcomings of the pioneering work on modelling diagnostic reasoning as a whole by Elstein et al. [24] back in 1978 led to a general discussion about the cognitive representation of medical knowledge and expertise in the mind of a trained physician. Improvement ideas gave rise to non-analytical reasoning processes like pattern recognition [25, p. 440] and recently proposed models of clinical reasoning acknowledge the presence of such heuristic shortcuts mostly framed as System 1 thinking [26]. Thus, there is a tight connection between proposed models of reasoning in the medical domain and Kahnemann's book on "Thinking, fast and slow" [27] up to the point that it inspired Coughlan et al. to publish a review on "Diagnosing, fast and slow" which ultimately concludes that "*Cognitive biases are ubiquitous, even among experts. Doctors do not differ in this regard*" [28, p. 5]. More detailed analyses have identified approximately 50 different biases affecting a physician during clinical reasoning [29]. Among these the ones which have been studied most extensively in the field of diagnostics are an availability bias [30, 14, 31, 32, 33, 34, 15], confirmation bias [14, 35, 15, 34], premature closure [14, 13, 15, 36], overconfidence [37, 14, 34, 35], anchoring bias [38, 15, 34, 36, 32] and a representativeness heuristic [39, 38, 15, 32, 34, 40].

Moving away from the descriptive and empirical level to examine cognitive biases, little research has been conducted on a formal decision-theoretic formalization of each one. In a resource allocation problem called police patrol hour assignment, Wu et al. introduce "*a general framework for deceiving adversaries with bounded rationality in terms of the obtained reward minimization*" [41, p. 6] by leveraging prospect theory [42] to compute a biased reward function of a human opponent. Then a MDP-based control policy is derived to exploit the differences between the environments objective and the adversary's subjective reward for maximal deception. Zhang et al. [43] apply deceptive kernel functions to observations emitted in a POMDP model to trick the agent in establishing false beliefs about the environment. And Bilinski et al. speculate on demonstrating bias in a cyber deception game by "*exploration of game parameters [...] to associate the outcome of algorithms to that of human decision-making biases*" [44]. More specifically, they suggest adaptations to their domain to potentially incorporate a gambler's or sunk cost fallacy.

Overall, there is a growing interest in understanding and measuring cognitive biases during medical diagnosis over recent decades. While research on this topic is extensive on the descriptive level, and rising on the empirical level, our approach is - to the best of our knowledge - the first to operationalize cognitive biases in a formal decision-theoretic framework like Markov decision processes with the goal to deliberately reproduce cognitive reasoning errors during medical diagnosis.

3. Modeling diagnostic reasoning

The idea behind the modeling approach is to capture the problem of medical diagnosis in a POMDP and then exploit the transparency in the formalism to integrate biased components

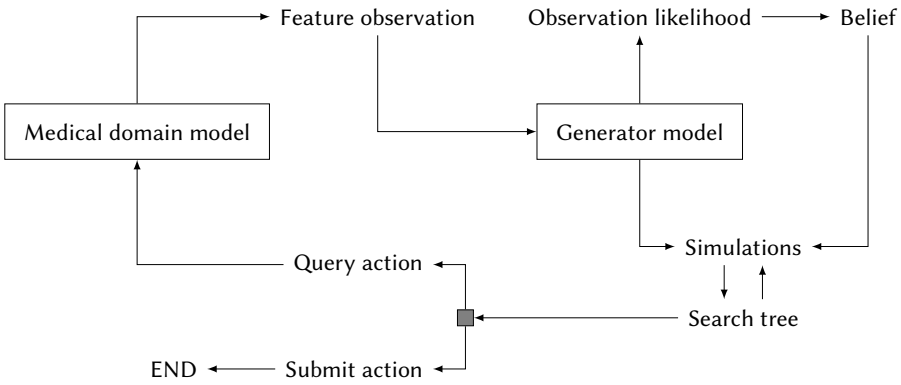

Figure 1: Schematic view on the diagnostic process model.

into the decision problem representation. As sketched in Figure 1, the agent can decide to query specific medical information which is then sampled from the medical domain model. The obtained information updates the belief through its likelihood to be observed given each diagnosis and the agent simulates possible next queries to choose the most suitable one at the current step in the process. This cycle of information retrieval repeats until the agent submits a diagnosis. Biases can manifest either in the retrospective evaluation of observations in the belief update or the prospective assessment of potential actions. As visualized in Figure 1 both cases are tied to the generator model. With this broad overview at hand, Section 3.1 sheds light on how the medical domain model is defined and used for sampling, Section 3.2 defines the resulting POMDP with all of its components and Section 3.3 elaborates on the modifications made to the generator model for each cognitive bias.

3.1. Medical domain model

We use a Bayesian network to model the medical domain stochastically. In general, Bayesian networks represent a joint probability distribution over a finite set of discrete random variables and enable flexible inference of prior or conditional distributions over node values [45]. Our network defines how medical features influence others, i.e., how diseases are caused and how diseases present themselves in form of symptoms. While there is a causal connection between disease and findings, observations might differ from patient to patient. It therefore makes sense to model the effect of a disease with a joint probability distribution over possible diagnoses and medical findings. The structure of the network is derived from the work of Richens et al. [46] who propose a three layer disease model where the top layer describes medical risk factors, the mid layer all diseases and the bottom layer potential symptoms. The nodes for risk factors and symptoms are derived from the work of Wardrope et al. [2]. They conducted a study on the prevalence of medical features among patients suffering from either epileptic seizures, syncopes or psychogenic non-epileptic seizures. Through an iterative random forest approach, they calculated that 36 features were enough to optimally predict the disease of a patient. Each of these features is adopted as one binary node in the network - except whether or not the patient had a brain tumor because the sample size was too low. Features concerning

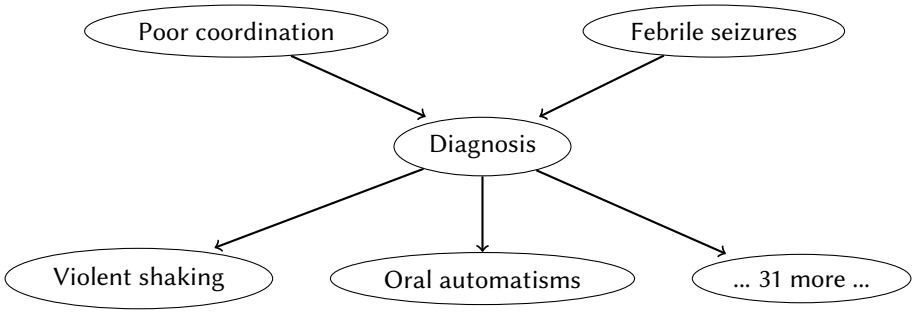

Figure 2: The Bayesian network used as medical domain model for the differential diagnosis between epileptic seizures, psychogenic non-epileptic seizures and syncopes. The structure is derived from [46], nodes and parameters are derived from [2]. All nodes are binary with values *yes* or *no* except for the diagnosis node that holds all three diseases.

the general medical history of the patient count as risk factor in the top layer, other features about the semiology of a specific seizure count as symptom in the bottom layer. Having suffered from febrile seizures in the childhood is thus considered a risk factor while violently shaking during a seizure is a symptom. The parameters of each node are calculated as a maximum likelihood estimate from the prevalence data presented in the paper [2]. This ensures, that the final medical domain model (see Figure 2) captures actual medical relations from the real world. With this model at hand, we can run the variable elimination algorithm [45] to flexibly compute the likelihood $P(f_{n+1}|d, f_1, \ldots, f_n)$ of any feature f_{n+1} given a specific diagnosis d and a set of already known features $f_1, \ldots, f_n$.

3.2. Diagnostic process model

We use a POMDP to model the diagnostic reasoning process. It is formally defined as the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, R, \gamma \rangle$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions and $\mathcal{O}$ a set of observations. T defines the transition model from one state to the next, Z defines the probability of emitting an observation given the current state and action and R likewise describes the probability of emitting reward given the current state and action. γ is a discount factor to decrease the influence of reward collected far in the future.

In our context, a diagnosis is understood as the sequential exploration of medical findings up to a point where the agent is sufficiently certain to commit itself to one of multiple options. The agent holds an initially uniform belief distribution over all three diagnostic options. Then, the agent is able to query for medical information (i.e., risk factors and symptoms in the medical domain model) and obtains uncertain observations (i.e., a sampled value for this node from the medical domain model). The crucial decision to make at each step is which medical feature to query next. Solving the POMDP for a diagnostic policy thus defines one trajectory through the space of available medical knowledge. After receiving the feature observation the agent updates its belief state depending on the likelihood of observing such a feature value in a patient given each diagnosis and the interaction loop starts again (see Figure 1). At any point in time, the agent can submit a specific diagnosis and by that end the whole process. Derived policies are therefore expected to exhibit a trade-off between gathering valuable information about

the patient to reduce uncertainty and keeping the diagnostic process as short and efficient as possible.

We solve the POMDP problem via partially observable upper confidence bounds for trees in Monte-Carlo planning (PO-UCT) [47] using the Python library `pomdp_py` [16]. This algorithm is essentially an adapted version of a Monte-Carlo tree search where state nodes in the tree are replaced by histories and the greedy action selection is adjusted for already encountered nodes in the tree. For unknown nodes, we use a uniform random action selection. At each time step, the agent runs simulations by sampling a current state from its belief state and then repeatedly using a generator model to simulate transitioning to state s' and obtaining observation o and reward r after executing action a in state s in each trajectory. After all simulations, the agent executes the action that reached the maximum expected return in the search tree. In the following, we will introduce each POMDP component in more detail.

State space Each state $s \in \mathcal{S}$ is defined by the true diagnosis the patient is suffering from (i.e., epileptic seizure, syncope or psychogenic non-epileptic seizure) and a predicted diagnosis by the agent. This predicted diagnosis is *null* as long as the process is running and is set accordingly once the agent commits itself. Thus, states where the true diagnosis and the predicted diagnosis are set are terminal states.

Action space Each action $a \in \mathcal{A}$ has one of two types. There are query actions to ask for specific information (one query action for each risk factor and symptom) and submit actions to commit oneself to one option (one for each disease). For example, executing *query febrile seizures* means to ask the patient whether he has ever suffered from febrile seizures.

Observation space Each observation $o \in \mathcal{O}$ is associated with a node in the medical domain model. After query actions, the observation holds the queried feature and a value of that feature. Observing *febrile seizures = no* thus means that the patient has not suffered from febrile seizures in the past. After submit actions, the observation just holds the submitted diagnosis as the process ends.

Transition model Query actions do not change the state. Submit actions lead to a deterministic transition into a terminal state, where the predicted diagnosis is set accordingly. Once the agent is in a terminal state, no transitions are possible.

Observation model The likelihood $P(o|s', a)$ of obtaining observation o after transitioning to state s' via query action a is computed using the medical domain model. The distribution over all values of the queried feature is conditioned on the true diagnosis d in the state and the medical evidence $o_1, \ldots, o_n$ obtained thus far from previous observations. The next feature value for observation o_{n+1} is then sampled from $P(o_{n+1}|d, o_1, \ldots, o_n)$. Observations after submit actions are deterministic.

Reward model Executing query actions triggers a reward of -1 to incentivise a short diagnostic trajectory that does not query more features than necessary. Submit actions are

rewarded with $+100$ for a correct diagnosis (i.e., true and predicted diagnosis in the state match) and -100 for a wrong one.

Generator model The generator model is used during simulations to generate a successor state s', an observation o and a reward r when mentally executing action a in state s. It reuses the observation model to infer the likelihood of a feature value $P(o|s', a)$ to sample from it. Query actions again lead to a reward of -1. As the agent cannot know the true diagnosis during simulation, the reward for submit actions differs from the true reward model. Here, reward is dependent on the entropy in the belief state. If the agent submits the diagnosis it is most convinced of and the normalized entropy in its belief state is below 0.25, then it collects a reward of $+100$ and -100 otherwise to penalize unjustified diagnoses under high uncertainty.

3.3. Modeling cognitive biases

We simulate cognitive biases with the diagnostic policy via adjustments to the generator model during planning with PO-UCT. As mentioned earlier, the agent determines the next action to execute at a given time step by running multiple simulations from the current belief state. The evolution of these simulated trajectories is determined by the generator model G. Now if the environment defines a ground truth transition model T, observation model Z and reward model R, then cognitive biases manifest themselves in conceptual deviations from T, Z and R in G. In a cognitive sense, the agent exhibits a biased policy because its understanding of the environment and thus its mentally derived action outcomes are biased. To exemplify the effects of biases in the generator model, we integrate three of the most common cognitive biases associated with medical diagnoses: an availability bias, a confirmation bias and premature closure (see Section 2). In the following, we will explain the integration of each in more detail.

Availability bias "*The disposition to judge things as being more likely, or frequently occurring, if they readily come to mind. Thus, recent experience with a disease may inflate the likelihood of its [sic] being diagnosed*" [15, p. 777]. Neurologists may overestimate the prevalence of epileptic seizures where clinical psychiatrists tend to see psychogenic non-epileptic seizures as both draw expertise from past experience with either disease. We implement this bias during belief updates by using a skewed observation likelihood in the generator model. Before applying the update, each likelihood is scaled according to how "readily the corresponding diagnosis hypothesis comes to mind". The bias always targets one of the three options. Observation likelihoods given this diagnosis are scaled by 0.4 and others are scaled by 0.3, expressing a constant preference towards one available diagnosis.

Confirmation bias "*This entails underestimating information that does not fit one's hypothesis*" [14, p. 3]. Because this bias again affects the integration of new evidence into the current belief state, it is implemented similarly to the availability bias. We again target the agent's belief update by skewing the likelihood of observations in the generator model, but this time the likelihood of an observation given a specific diagnosis is not scaled by a fixed amount, but by the current belief of the agent in that particular diagnosis. Thus, it is harder for new evidence to

Bias	Mean trajectory length
None	17.8 ± 12.5
Availability bias	17.0 ± 10.5
Confirmation bias	5.3 ± 1.1
Premature closure	5.5 ± 2.6

Table 1
The length of the diagnostic trajectories (i.e., the number of queried features) for all agent variants.

alter the current belief states. Confirming observations are over- and contradictory observations are underestimated.

Premature closure "*The tendency to apply premature closure [...] accepting a diagnosis before it has been fully verified*" [15, p. 778] This bias manifests in the process if the agent commits itself to one diagnosis although the entropy in the belief state is too high to justify that. Consequently, we implement this bias by increasing the normalized entropy threshold in the generator model at which the agent is able to collect positive reward for a diagnosis submission from 0.25 to 0.75. This leads to the submission of unverified, uncertain diagnoses early in the process.

4. Experiments

To demonstrate the resulting policies under each bias, we run and track 30 diagnostic trajectories (10 for each differential diagnosis) for each possible variant of the agent: unbiased, availability biased towards epileptic seizures, availability biased towards psychogenic seizures, availability biased towards syncopes, confirmation biased and affected by premature closure. For computing the next action to execute, we run 200 simulations at each time step via PO-UCT planning with a maximum depth of 50 during tree search, an exploration constant of 1.0 and a discount factor of 0.9.

Unbiased Among all variants, the unbiased agent achieves a diagnostic accuracy of 86.7% (see Table 2) which is comparable to the 86.0% reported by Wardrope et al. [2] on the same data set that has been used to learn the parameters of the medical domain model. During diagnosis, the agent exhibits no particular preference for specific risk factors or symptoms (see Figure 3) and revises its diagnostic belief over time by changing its main hypothesis (see Figure 4).

Availability bias In contrast to unbiased policies, agents with an availability bias focus on a subset of features during knowledge exploration (see Figure 3). While the length of each diagnostic process is comparably long to not having any bias (see Table 1), a previous familiarity with epileptic seizures leads to a focus on limp limbs and oral automatisms. A similar effect is observable for psychogenic seizures (focused on deja-vu and impaired awareness) and syncopes (focused on poor coordination and rapid head turning) but the effect remains

strongest for epileptic seizures. Concerning diagnostic performance, an agent primed towards epileptic seizures misdiagnoses nearly all cases of psychogenic seizure (9/10) and vice versa the agent primed towards psychogenic seizures even declares all cases of epileptic seizures as psychogenic. This effect is absent for the diagnosis of syncopes which may be caused by the closer similarity between epileptic and psychogenic seizures. Wardrope et al. [2] report a similar effect concerning confusion matrices between all three diagnoses, where all syncopes are identified correctly while epileptic and psychogenic seizures are mixed up more often.

Bias	ES	PNES	S	Accuracy
None	**90.0%**	**70.0%**	**100.0%**	**86.7%**
Availability bias	63.3%	63.3%	86.7%	71.1%
Confirmation bias	20.0%	**70.0%**	80.0%	56.7%
Premature closure	70.0%	**70.0%**	90.0%	76.7%

Table 2
The number of correctly identified seizures grouped by epileptic seizures (ES), psychogenic seizures (PNES), syncopes (S), and the overall diagnostic accuracy of all agent variants.

Confirmation bias A confirmation bias leads to short and unsuccessful diagnostic processes (see Table 1 and 2). The agent typically follows its first main hypothesis and after a few steps confirms itself more and more into wrong diagnoses. As opposed to the unbiased model, the policy is not able to recover from misleading belief states, that would need to be reverted upon contradictory information (see Figure 4).

Premature closure Similar to a confirmation bias, trajectories affected by premature closure are very short. But resulting policies still perform remarkably well in terms of accuracy (see Table 2). In a vital difference to the confirmation bias, the diagnostic process is therefore not impaired at the stage of information integration, the agent simply does not collect enough information to integrate.

5. Conclusion

In conclusion, this paper proposes that cognitive biases are rational behaviour in misunderstood environments. Formalizing a cognitive bias during diagnostic reasoning means distorting the true environment to render the biased policy as optimal policy. We achieve these distortions by skewing rewards and observation likelihoods during belief updates, but any modification to the environment would be viable in principle. Agents could neglect parts of the state or action space or dynamically grow the proportions of action they know. We justify this perspective by the fact that planning in itself always requires some mechanism for mental outcome prediction. If these outcome predictions are based on wrong assumptions, planning will produce biased strategies.

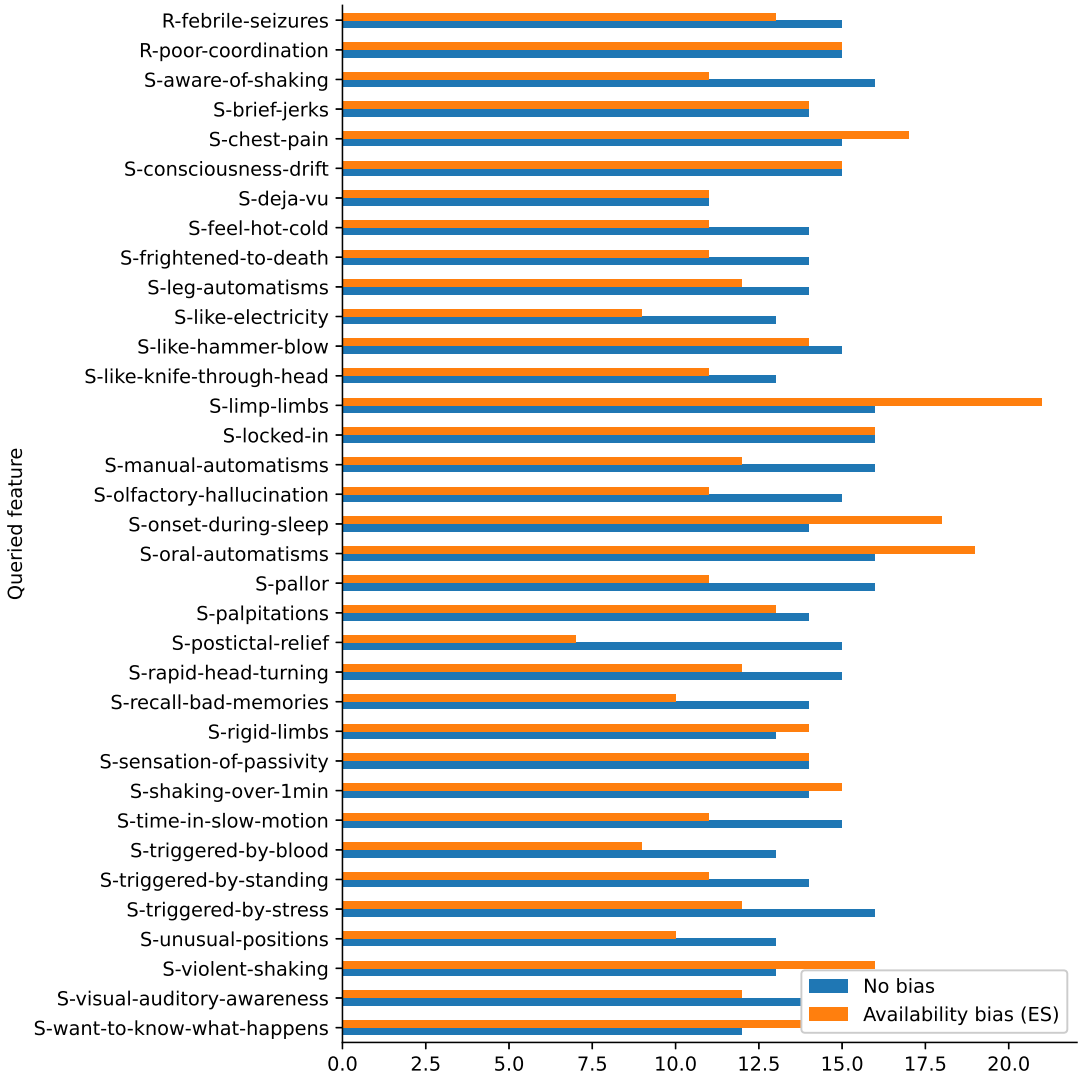

Figure 3: Number of queries by feature in 30 different diagnostic trajectories. Blue bars indicate the number of trajectories in which a certain feature was queried by the unbiased model, likewise orange bars indicate the same for the model that was availability biased towards epileptic seizures. The unbiased model queries each feature rather uniformly between 12 - 16 times, while the biased model shows selective preference towards limp limbs and oral automatisms while neglecting postictal relief.

As we have shown, partially observable Markov decision processes in combination with Monte-Carlo planning methods have proven as a suitable framework to implement these ideas in an explainable fashion. By manipulating the generator model, biases due to skewed Bayesian reasoning can be explicitly represented. Agents with an availability bias are deceived into misdiagnoses of similar diseases (epileptic vs. psychogenic seizures), confirmation biased physicians tend to arrive at a fast and mostly wrong diagnosis and premature closure can

be induced by rewarding prematurely. But the approach in general is at a rather early stage and will need much more investigation. How this approach relates to modeling the proposed System 1 and 2 processes during diagnosis [28] explicitly will require further research especially with regard to cognitive plausibility. At the moment, a simplistic Bayesian network as medical domain model is used, where only 35 features are integrated and each feature can only be present or absent. The patient always suffers from one of three diseases. We plan to enhance the current approach with an empirical study to collect diagnostic trajectories from physicians. Analyzing their queries and diagnostic strategies will enable improvements to both the medical domain model and the process model. Overall, the presented work contributes to empowering diagnostic reasoning support systems to detect and potentially classify erroneous diagnostic reasoning. As we can only fix what we understand, identifying the error cause is a first important step towards resolving misdiagnoses altogether.

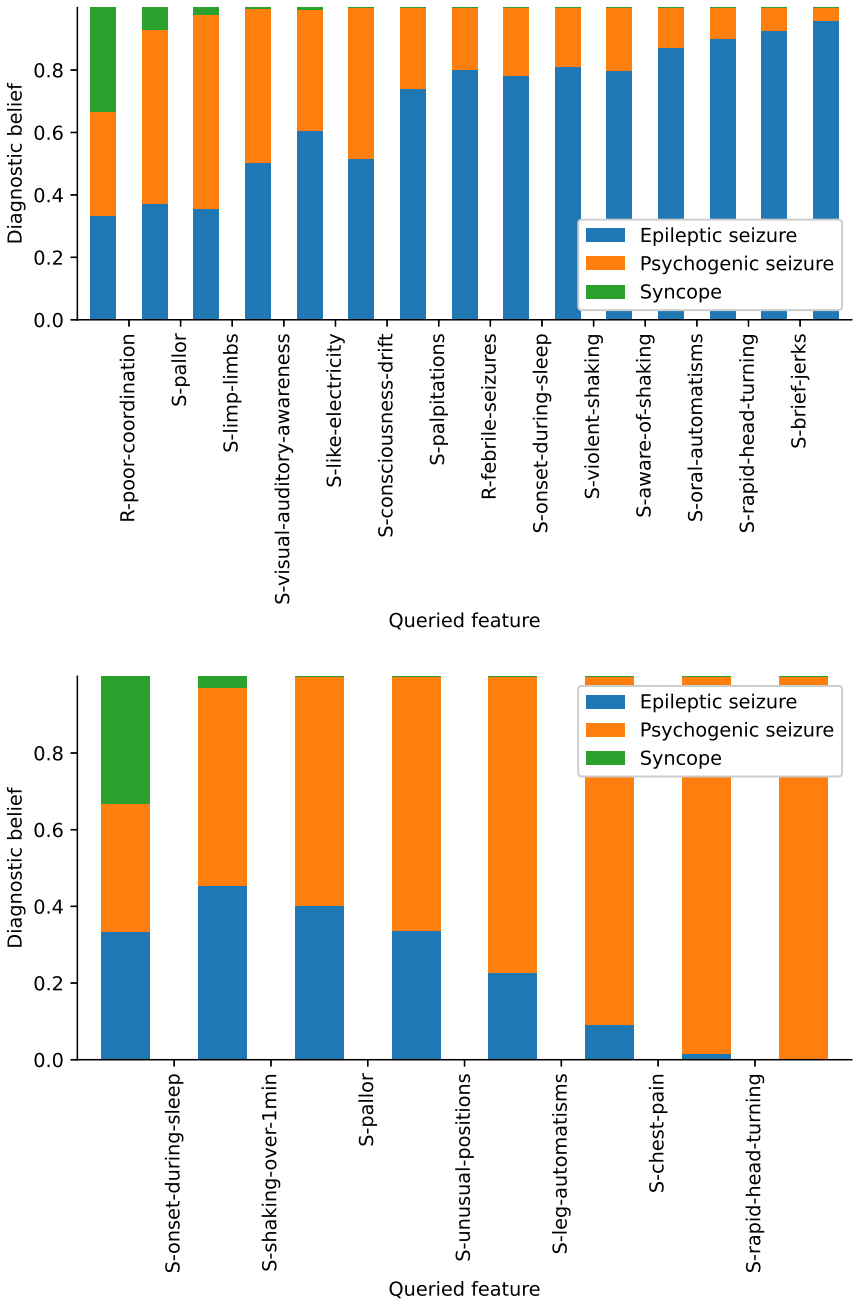

Figure 4: Diagnostic belief change over time for the unbiased model (top) and the confirmation biased model (bottom) for a patient suffering from an epileptic seizure. Each bar composed of blue, orange and green represents the belief state distribution at one time step. Time (and thus the belief) evolves over the x axis with each new feature queried. While the unbiased model can recover from a state where a psychogenic seizure looked more probable (top, step 3), the confirmation biased model simply confirms itself more and more into the wrong diagnosis.

Acknowledgments

Funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

[1] E. Foster, P. Carney, D. Liew, Z. Ademi, T. O'Brien, P. Kwan, First seizure presentations in adults: beyond assessment and treatment, Journal of Neurology, Neurosurgery & Psychiatry 90 (2019) 1039–1045. URL: https://jnnp.bmj.com/lookup/doi/10.1136/jnnp-2018-320215. doi:10.1136/jnnp-2018-320215.

[2] A. Wardrope, J. Jamnadas-Khoda, M. Broadhurst, R. A. Grünewald, T. J. Heaton, S. J. Howell, M. Koepp, S. W. Parry, S. Sisodiya, M. C. Walker, M. Reuber, Machine learning as a diagnostic decision aid for patients with transient loss of consciousness, Neurology: Clinical Practice 10 (2020) 96–105. URL: https://cp.neurology.org/lookup/doi/10.1212/CPJ.0000000000000726. doi:10.1212/CPJ.0000000000000726.

[3] M. M. Oto, The misdiagnosis of epilepsy: Appraising risks and managing uncertainty, Seizure 44 (2017) 143–146. URL: https://linkinghub.elsevier.com/retrieve/pii/S1059131116302977. doi:10.1016/j.seizure.2016.11.029.

[4] K. Malmgren, M. Reuber, R. Appleton, Differential diagnosis of epilepsy, Oxford textbook of epilepsy and epileptic seizures (2012) 81–94.

[5] C. A. van Donselaar, H. Stroink, W.-F. Arts, for the Dutch Study Group of Epilepsy in Childhood, How Confident Are We of the Diagnosis of Epilepsy?, Epilepsia 47 (2006) 9–13. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1528-1167.2006.00653.x. doi:10.1111/j.1528-1167.2006.00653.x.

[6] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, npj Digital Medicine 3 (2020) 17. URL: http://www.nature.com/articles/s41746-020-0221-y. doi:10.1038/s41746-020-0221-y.

[7] E. S. Berner, Clinical decision support systems: theory and practice, 2007. URL: https://www.researchgate.net/profile/J-Marc-Overhage/publication/227221582_Clinical_Decision_Support_Within_the_Regenstrief_Medical_Record_System/links/0deec51db21308243e000000/Clinical-Decision-Support-Within-the-Regenstrief-Medical-Record-System.pdf.

[8] J. P. Zöllner, S. Wolking, Y. Weber, F. Rosenow, Decision-support-Systeme, Assistenzsysteme und Telemedizin in der Epileptologie, Der Nervenarzt 92 (2021) 95–106. URL: http://link.springer.com/10.1007/s00115-020-01031-7. doi:10.1007/s00115-020-01031-7.

[9] W. F. Bond, L. M. Schwartz, K. R. Weaver, D. Levick, M. Giuliano, M. L. Graber, Differential Diagnosis Generators: an Evaluation of Currently Available Computer Programs, Journal of General Internal Medicine 27 (2012) 213–219. URL: http://link.springer.com/10.1007/s11606-011-1804-8. doi:10.1007/s11606-011-1804-8.

[10] V. Patterson, P. Pant, N. Gautam, A. Bhandari, A Bayesian tool for epilepsy diagnosis in the resource-poor world: Development and early validation, Seizure 23 (2014) 567–

569. URL: https://linkinghub.elsevier.com/retrieve/pii/S1059131114000922. doi:10.1016/j.seizure.2014.03.010.

[11] T. Cook, J. C. Gee, R. N. Bryan, J. T. Duda, P.-H. Chen, E. Botzolakis, S. Mohan, A. Rauschecker, J. Rudie, I. Nasrallah, Bayesian network interface for assisting radiology interpretation and education, in: J. Zhang, P.-H. Chen (Eds.), Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, SPIE, Houston, United States, 2018, p. 26. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10579/2293691/Bayesian-network-interface-for-assisting-radiology-interpretation-and-education/10.1117/12.2293691.full. doi:10.1117/12.2293691.

[12] S. Baalen, M. Boon, P. Verhoef, From clinical decision support to clinical reasoning support systems, Journal of Evaluation in Clinical Practice 27 (2021) 520–528. URL: https://onlinelibrary.wiley.com/doi/10.1111/jep.13541. doi:10.1111/jep.13541.

[13] M. L. Graber, N. Franklin, R. Gordon, Diagnostic Error in Internal Medicine, Archives of Internal Medicine 165 (2005) 1493. URL: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.165.13.1493. doi:10.1001/archinte.165.13.1493.

[14] T. Watari, Y. Tokuda, Y. Amano, K. Onigata, H. Kanda, Cognitive Bias and Diagnostic Errors among Physicians in Japan: A Self-Reflection Survey, International Journal of Environmental Research and Public Health 19 (2022) 4645. URL: https://www.mdpi.com/1660-4601/19/8/4645. doi:10.3390/ijerph19084645.

[15] P. Croskerry, The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them, Academic Medicine 78 (2003) 775–780. URL: http://journals.lww.com/00001888-200308000-00003. doi:10.1097/00001888-200308000-00003.

[16] K. Zheng, S. Tellex, pomdp_py: A Framework to Build and Solve POMDP Problems, 2020. URL: http://arxiv.org/abs/2004.10099, arXiv:2004.10099 [cs].

[17] W. Zhang, H. Wang, Diagnostic Policies Optimization for Chronic Diseases Based on POMDP Model, Healthcare 10 (2022) 283. URL: https://www.mdpi.com/2227-9032/10/2/283. doi:10.3390/healthcare10020283.

[18] J. Zhang, B. T. Denton, H. Balasubramanian, N. D. Shah, B. A. Inman, Optimization of Prostate Biopsy Referral Decisions, Manufacturing & Service Operations Management 14 (2012) 529–547. URL: http://pubsonline.informs.org/doi/abs/10.1287/msom.1120.0388. doi:10.1287/msom.1120.0388.

[19] Y. Li, M. Zhu, R. Klein, N. Kong, Using a partially observable Markov chain model to assess colonoscopy screening strategies – A cohort study, European Journal of Operational Research 238 (2014) 313–326. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221714002185. doi:10.1016/j.ejor.2014.03.004.

[20] E. F. Arruda, B. B. Pereira, C. A. Thiers, B. R. Tura, Optimal testing policies for diagnosing patients with intermediary probability of disease, Artificial Intelligence in Medicine 97 (2019) 89–97. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365717301513. doi:10.1016/j.artmed.2018.11.005.

[21] N. Bazrafshan, M. M. Lotfi, A finite-horizon Markov decision process model for cancer chemotherapy treatment planning: an application to sequential treatment decision making in clinical trials, Annals of Operations Research 295 (2020) 483–502. URL: https://link.springer.com/10.1007/s10479-020-03706-5. doi:10.1007/s10479-020-03706-5.

[22] M. Hauskrecht, H. Fraser, Planning treatment of ischemic heart disease with partially observable Markov decision processes, Artificial Intelligence in Medicine 18 (2000) 221–244. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365799000421. doi:10.1016/S0933-3657(99)00042-1.

[23] R. Ibrahim, B. Kucukyazici, V. Verter, M. Gendreau, M. Blostein, Designing Personalized Treatment: An Application to Anticoagulation Therapy, Production and Operations Management 25 (2016) 902–918. URL: https://onlinelibrary.wiley.com/doi/10.1111/poms.12514. doi:10.1111/poms.12514.

[24] A. S. Elstein, L. S. Shulman, S. H. Sprafka, S. A. Sprafka, Medical problem solving an analysis of clinical reasoning (1978).

[25] C. Koufidis, K. Manninen, J. Nieminen, M. Wohlin, C. Silén, Unravelling the polyphony in clinical reasoning research in medical education, Journal of Evaluation in Clinical Practice 27 (2021) 438–450. URL: https://onlinelibrary.wiley.com/doi/10.1111/jep.13432. doi:10.1111/jep.13432.

[26] P. Croskerry, A Universal Model of Diagnostic Reasoning, Academic Medicine 84 (2009) 1022–1028. URL: http://journals.lww.com/00001888-200908000-00014. doi:10.1097/ACM.0b013e3181ace703.

[27] D. Kahneman, Thinking, fast and slow, Macmillan, 2011.

[28] J. Coughlan, C. F. Mullins, T. J. Kiernan, Diagnosing, fast and slow, Postgraduate Medical Journal 97 (2021) 103–109. URL: https://pmj.bmj.com/lookup/doi/10.1136/postgradmedj-2019-137412. doi:10.1136/postgradmedj-2019-137412.

[29] P. Croskerry, 50 Cognitive and Affective Biases in Medicine, 2013. URL: https://sjrhem.ca/wp-content/uploads/2015/11/CriticaThinking-Listof50-biases.pdf.

[30] P. Li, Z. y. Cheng, G. l. Liu, Availability Bias Causes Misdiagnoses by Physicians: Direct Evidence from a Randomized Controlled Trial, Internal Medicine 59 (2020) 3141–3146. URL: https://www.jstage.jst.go.jp/article/internalmedicine/59/24/59_4664-20/_article. doi:10.2169/internalmedicine.4664-20.

[31] S. Mamede, T. van Gog, K. van den Berge, R. M. J. P. Rikers, J. L. C. M. van Saase, C. van Guldener, H. G. Schmidt, Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy Among Internal Medicine Residents, JAMA 304 (2010) 1198. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2010.1276. doi:10.1001/jama.2010.1276.

[32] M. Richie, S. A. Josephson, Quantifying Heuristic Bias: Anchoring, Availability, and Representativeness, Teaching and Learning in Medicine 30 (2018) 67–75. URL: https://www.tandfonline.com/doi/full/10.1080/10401334.2017.1332631. doi:10.1080/10401334.2017.1332631.

[33] S. Monteiro, J. Sherbino, J. S. Ilgen, E. M. Hayden, E. Howey, G. Norman, The effect of prior experience on diagnostic reasoning: exploration of availability bias, Diagnosis 7 (2020) 265–272. URL: https://www.degruyter.com/document/doi/10.1515/dx-2019-0091/html. doi:10.1515/dx-2019-0091.

[34] J. S. Blumenthal-Barby, H. Krieger, Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy, Medical Decision Making 35 (2015) 539–557. URL: http://journals.sagepub.com/doi/10.1177/0272989X14547740. doi:10.1177/0272989X14547740.

[35] D. F. Whelehan, K. C. Conlon, P. F. Ridgway, Medicine and heuristics: cognitive biases and medical decision-making, Irish Journal of Medical Science (1971 -) 189 (2020) 1477–1484. URL: https://link.springer.com/10.1007/s11845-020-02235-1. doi:10.1007/s11845-020-02235-1.

[36] M. Gäbler, Denkfehler bei diagnostischen Entscheidungen, Wiener Medizinische Wochenschrift 167 (2017) 333–342. URL: http://link.springer.com/10.1007/s10354-017-0570-6. doi:10.1007/s10354-017-0570-6.

[37] E. S. Berner, M. L. Graber, Overconfidence as a Cause of Diagnostic Error in Medicine, The American Journal of Medicine 121 (2008) S2–S23. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002934308000405. doi:10.1016/j.amjmed.2008.01.001.

[38] B. G. Vickrey, M. A. Samuels, A. H. Ropper, How neurologists think: A cognitive psychology perspective on missed diagnoses, Annals of Neurology 67 (2010) 425–433. URL: https://onlinelibrary.wiley.com/doi/10.1002/ana.21907. doi:10.1002/ana.21907.

[39] G. R. Corazza, M. V. Lenti, P. D. Howdle, Diagnostic reasoning in internal medicine: a practical reappraisal, Internal and Emergency Medicine 16 (2021) 273–279. URL: http://link.springer.com/10.1007/s11739-020-02580-0. doi:10.1007/s11739-020-02580-0.

[40] G. Saposnik, D. Redelmeier, C. C. Ruff, P. N. Tobler, Cognitive biases associated with medical decisions: a systematic review, BMC Medical Informatics and Decision Making 16 (2016) 138. URL: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0377-1. doi:10.1186/s12911-016-0377-1.

[41] B. Wu, M. Cubuktepe, S. Bharadwaj, U. Topcu, Reward-Based Deception with Cognitive Bias, in: 2019 IEEE 58th Conference on Decision and Control (CDC), IEEE, Nice, France, 2019, pp. 2265–2270. URL: https://ieeexplore.ieee.org/document/9029476/. doi:10.1109/CDC40024.2019.9029476.

[42] D. Kahneman, A. Tversky, Prospect Theory: An Analysis of Decision under Risk, Econometrica 47 (1979) 263. URL: https://www.jstor.org/stable/1914185?origin=crossref. doi:10.2307/1914185.

[43] Z. Zhang, Q. Zhu, Deceptive Kernel Function on Observations of Discrete POMDP, 2020. URL: http://arxiv.org/abs/2008.05585, arXiv:2008.05585 [cs].

[44] M. Bilinski, J. diVita, K. Ferguson-Walter, S. Fugate, R. Gabrys, J. Mauger, B. Souza, Lie Another Day: Demonstrating Bias in a Multi-round Cyber Deception Game of Questionable Veracity, in: Q. Zhu, J. S. Baras, R. Poovendran, J. Chen (Eds.), Decision and Game Theory for Security, volume 12513, Springer International Publishing, Cham, 2020, pp. 80–100. URL: https://link.springer.com/10.1007/978-3-030-64793-3_5. doi:10.1007/978-3-030-64793-3_5, series Title: Lecture Notes in Computer Science.

[45] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[46] J. G. Richens, C. M. Lee, S. Johri, Improving the accuracy of medical diagnosis with causal machine learning, Nature Communications 11 (2020) 3923. URL: https://www.nature.com/articles/s41467-020-17419-7. doi:10.1038/s41467-020-17419-7.

[47] D. Silver, J. Veness, Monte-Carlo Planning in Large POMDPs, in: J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems, volume 23, Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/file/edfbe1afcf9246bb0d40eb4d8027d90f-Paper.pdf.