Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230946

Evaluating Methods for Setting a Prior Probability of Guilt

Ludi VAN LEEUWEN [a,1], Bart VERHEIJ [a], Rineke VERBRUGGE [a] and Silja RENOOIJ [b]

[a] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen*

[b] *Department of Information and Computing Sciences, Utrecht University*

Abstract. One way of reasoning with uncertainties in the context of law is to use probabilities. However, methods for reasoning about the probability of guilt in a court case requires us to specify a prior probability of guilt, which is the probability of guilt before any evidence is known. There is no accepted approach for specifying the prior probability of guilt but multiple solutions have been proposed. In this paper, we consider three approaches: a prior that is based on the population, a prior based on the number of agents that have similar opportunity as the suspect and a prior that represents a legal norm. For comparing and evaluating the approaches, we use an agent-based model as a ground truth in which all probabilities are known. With the data generated in the ground truth model, we investigate how the choice of prior influences the posterior probability of guilt for both guilty and innocent agents. Using a decision threshold, we can determine the effect of the three approaches on the rates of correct and incorrect convictions and acquittals. We find that the opportunity prior results in higher rates of both correct convictions and false convictions and requires more assumptions and access to data and knowledge than the legal prior and population prior.

Keywords. Opportunity prior, Legal probabilism, Bayesian Networks, Agent-based modelling

1. Introduction

Judges reason under uncertainty when they decide in court cases. They weigh pieces of evidence to conclude whether a suspect is guilty. However, evidence presented in court might be difficult to interpret, incomplete or untrustworthy and how strongly any piece of evidence relates to the final verdict is often left implicit. The uncertainty in going from evidence to conclusions can be expressed explicitly through probabilistic degrees of belief. Reasoning about probabilities of events, for example through Bayes' Theorem, requires specifying prior and conditional probability distributions over the events. Hence, if the judge has to decide on a posterior probability of guilt given the evidence, the prior probability of guilt has to be assigned. This is the so-called 'problem of the prior' [1].

For some events, prior probabilities can be set using statistical estimates such as frequentist base rates. However, taking the base rate of guilt over all court cases results in

[1]Corresponding Author: Ludi van Leeuwen, l.s.van.leeuwen@rug.nl

a prior that is inconsistent with the presumption of innocence. For example, 90% of the court cases in The Netherlands end in conviction,[2] therefore, only weak evidence would be necessary to come to a posterior probability of guilt that is above the probability of 0.95 that is typically proposed as the threshold of 'beyond reasonable doubt' [1].

As an alternative to the conviction base rate, other approaches have been proposed for establishing a prior probability of guilt by Lindley [2], Fenton, Neil, Dahlman and Lagnado [3] and Dahlman [4]. Both Lindley and Fenton et al. propose a prior that is based on location and population of a given crime scene. In contrast, Dahlman proposes a method for setting the prior of guilt based on a legal norm.

These different methods for setting priors have been proposed in the literature, but it is unclear how to compare and evaluate their performance. This paper proposes a method for testing the different methods of setting the priors by using an agent-based model (ABM) in combination with a Bayesian network (BN). In the ABM, criminal behaviour of agents can be simulated and the statistical patterns of behaviour by the agents can be represented in a Bayesian network. By setting different priors in the Bayesian network, the effects of each prior setting method on the false conviction and false acquittal rates can be investigated by using the simulation as a ground truth.

The rest of the paper is structured as follows. In Section 2, the three methods for setting the priors and their introduction into Bayesian Networks are introduced. In Section 3, the simulation and Bayesian Networks are described. Section 4 shows the results of the true and false positive rates of the networks given the priors used. Section 5 discusses existing problems in the literature in light of these findings. Section 6 concludes.

2. State of the Art

In this section, the different methods for setting the priors, which are the population prior, the legal prior and the opportunity prior are described. These priors will be used as priors in Bayesian Network idioms for reasoning. Bayesian Networks are introduced as a method for modelling the evidence in the simulation.

2.1. Methods for Setting Prior Probabilities of Guilt

Both the population prior and the opportunity prior assume that a crime has occurred, and the prior specifies who could have done it [4]. They take as a prior $\frac{1}{S}$, where S is the number of possible perpetrators.

The *Population* prior was proposed by Lindley ([2], p 218). The number of possible perpetrators S is set to the size of the population of the country in which the crime is committed.

The *Opportunity* prior was proposed by Fenton et al., this prior aims to set S based on the number of agents with a similar opportunity as a given suspect. This is formalized as follows: First the number of agents at the crime scene n is determined. One of these agents must be the perpetrator, hence, a prior probability of guilt of $\frac{1}{n}$. Second, opportunity is defined as the number of agents N who had the same opportunity as the suspect to be present at the crime scene. Since there are n agents at the crime scene and there are

[2]In 2021, in the Netherlands, in 90% of the court cases before appeal, the suspect was deemed guilty https://www.cbs.nl/nl-nl/cijfers/detail/83944NED.

N agents who could have been there with the same possibility as the suspect, the prior for being present at the crime scene is $\frac{n}{N}$. The number of agents who could have been present is based on the last known location of the given suspect. This way opportunity prior allows us to set the prior probabilities of two events: the event for a suspect being at the crime scene: $\frac{n}{N}$, and the event of committing the crime, conditioned on the event that the suspect was at the crime scene: $\frac{1}{n}$.

The *Legal* prior was proposed by Dahlman [4] and represents a legal norm. Instead of being based on a set of possible perpetrators, using this prior neither assumes that a crime has been committed, nor assumes our knowledge of the size of the set of possible perpetrators. Instead, this method conceptualizes the prior as a legal norm. Any case should start with a prior probability of guilt for any suspect of $\frac{1}{100}$ and a prior probability for not guilty of $\frac{99}{100}$, although these values are essentially arbitrary.

2.2. Bayesian Networks

A Bayesian Network (BN) is a tuple $\langle V, E, P \rangle$ consisting of a directed acyclic graph (DAG) with nodes V representing random variables and directed edges E, which together capture the independence relation over the variables in joint probability distribution P [5]. P is defined through the chain rule $P = \prod_{i=1}^{n} \Pr(V_i|parents(V_i))$. It combines for every node $V_i \in V$ the distributions $\Pr(V_i|parents(V_i))$ for every combination of values for the parents[3] of V_i in the graph; these distributions are specified in node probability tables (NPTs). We assume variables are Boolean with possible values $\{T, F\}$. BNs, or sub-structures of BNs called idioms, have been proposed for representing reasoning in court cases [6,7,8,9,10,11,12]. Fenton et al. proposed a BN idiom for reasoning about the opportunity prior [3], of which a simplified version is used in this paper (Figure 2, with the nodes *AtCrimeScene* and *Guilty*). Combined with a threshold t, BNs can work as classifiers. We will take a posterior probability of *Guilty* = *true* given the evidence above the threshold $t = 0.95$, i.e., $P(guilty|evidence) > 0.95$, to result in a positive, or 'guilty' classification, while a probability below the threshold would be a negative, or 'innocent' classification. The threshold of 0.95 was chosen because it expresses the concept 'beyond reasonable doubt', for the sake of the current discussion [1].

2.3. Agent-Based Models

Agent-Based Models (ABMs) have been used to provide a ground truth of data for testing methods in the legal domain [8]. They provide a controlled environment in which human-like behaviour can arise due to agent-agent and agent-world interactions. The data gathered from the simulation about the statistical trends present in the behaviour of the agents can then be used to model the simulation in a Bayesian Network.

3. Method

We used an ABM to model a simplified theft scenario where we know for sure that a theft was committed, by exactly one criminal. We use three pieces of evidence. In the ABM, agents with different roles moved around in a spatial environment and their behaviour was observed by us and modelled in a Bayesian Network.

[3]Parents of V_i are nodes V_j such that $V_j \rightarrow V_i$.

3.1. Simulation of the Theft Scenario

The ABM was programmed in a 25-by-25 patches Netlogo environment [13] in which 200 agents moved around between 8 stores. One of those 200 agents was the victim with a wallet, and there was one thief who attempted to steal the wallet.[4] All 8 stores had a camera that observed locations of the agents within a radius of 25 and an angle of 145, which were initialised randomly for every run. In every run, one agent was the victim with the wallet and there were three suspects: the thief, or the **guilty suspect**, an agent that was picked randomly out of the population: the **innocent random suspect** and the innocent agent that is the closest to the crime scene: the **innocent bystander suspect**. While both innocent suspects had the same behaviour as all the other agents, the thief had a special behavioural pattern: It had a vision range of 15 and a vision angle of 360, it saw a victim, moved towards them and stole the wallet. During the stealing, the thief could be caught red-handed. The Netlogo patch, or location at which the wallet was stolen, is the crime scene (CS) and the epoch at which it was stolen is the crime time (CT). After stealing, the thief would drop the wallet to get rid of the evidence. Other agents were able to pick up the wallet. The simulation ran on after the stealing event took place for on average 10 epochs.

There were three pieces of evidence in the simulation, which each could be true or false for each suspect. These pieces were: $EredHanded$, representing whether the suspect was caught red-handed during the crime; $Ewallet$, representing whether the suspect was holding the wallet at the end of the simulation; and *EseenCS*, representing whether the suspect was seen by cameras at CS at CT (T), seen away from CS (F), or not seen by cameras at all (*NA*).

For the opportunity prior, we counted the number of agents that were present on the crime-scene patch as N, and we calculated N by counting the number of agents that were within the area A in time Δt, where A is centered on CS, with radius r the distance between CS and a known location of the suspect at time x, and Δt was the time interval either between x and the crime time, or between the crime time and x.

The simulation was run 1000 times. In every run, information was collected to calculate n, N, which were the location and time of the crime, and the locations of all agents at every time step and the frequencies of $Guilty$, $EredHanded$, $Ewallet$, *EseenCS*. This is the ground truth. Additionally, in every run, for every suspect, an evidence set was collected: $\langle\, Ewallet = \{T,F\}, Eredhanded = \{T,F\}, EseenCS = \{T,F\}\,\rangle$, or when the agent is not seen on camera at CT $\langle\, Ewallet = \{T,F\}, Eredhanded = \{T,F\}\,\rangle$.

3.2. Modelling the Theft Scenario with Bayesian Networks

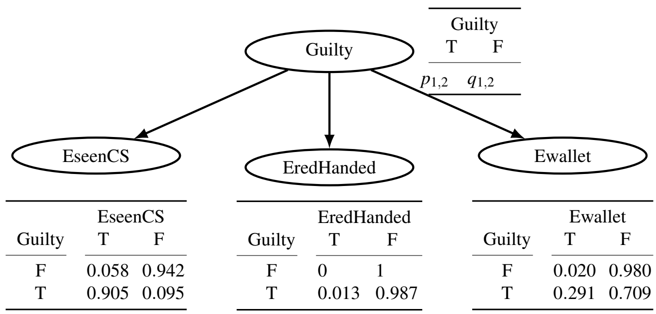
We created a model of the different priors as well as the events and evidence of the simulation, in Bayesian Networks. The nodes in the BN were as follows: For the evidence, we represented $Ewallet, EredHanded, EcameraCS$ as nodes. We represented the hypothesis of guilt of the agents in the node $Guilty$. For the opportunity prior method, we needed a node to account for the event of some agent being at the crime scene, for which we used the node *AtCrimeScene*.

The structure of the Bayesian Network depended on which prior we used. The legal and population BNs had the structure as shown in Figure 1, and connected each piece

[4] Simulation code and more extended description is available at https://github.com/aludi/JURIX2023

of evidence as a child to the *Guilty* node, following the evidence-idiom [10]. For the opportunity prior (Figure 2), we followed a simplified version of the opportunity idiom suggested by [3]. The structure was simplified because we did not have separate nodes for n and N, but instead calculated the values of the priors directly in the nodes.

The node probability tables (NPTs) shown for every BN are obtained from the frequencies of the events within the simulation over 1000 runs, except the parameters p_1, p_2, n, N. For the legal prior, we set $\Pr(guilty) = p_1 = \frac{1}{100}$ in the legal-BN. In the population prior we set $\Pr(guilty) = p_2 = \frac{1}{200}$ in the population-BN. $\Pr(guilty) = \frac{1}{n}$ and $\Pr(atCrimeScene) = \frac{n}{N}$. Since n and N differ per run of simulation and per suspect, a new BN based on Figure 2 was created for every run for every suspect, resulting in $1000 \cdot 3 = 3000$ BNs with each their own values of n and N.



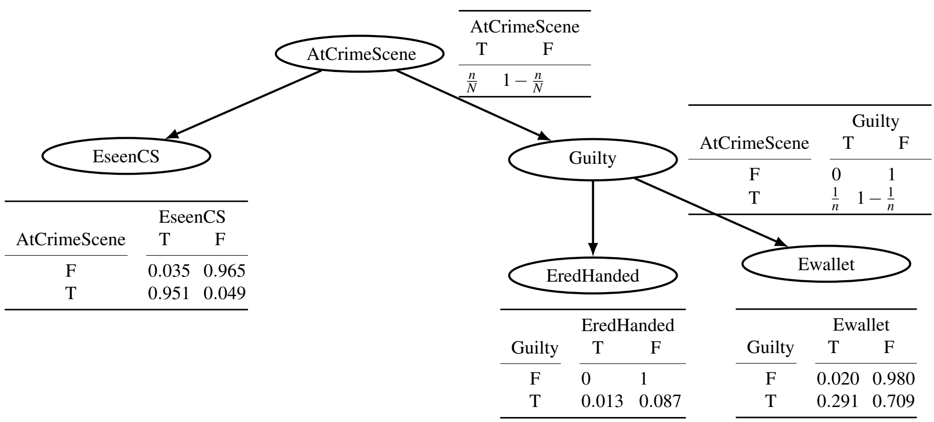
Guilty T	Guilty F
$p_{1,2}$	$q_{1,2}$

Guilty	EseenCS T	EseenCS F
F	0.058	0.942
T	0.905	0.095

Guilty	EredHanded T	EredHanded F
F	0	1
T	0.013	0.987

Guilty	Ewallet T	Ewallet F
F	0.020	0.980
T	0.291	0.709

Figure 1. Bayesian Network for Legal prior ($p_1 = \frac{1}{100}, q_1 = 1 - p_1$) and Population prior ($p_2 = \frac{1}{200}, q_2 = 1 - p_2$), legal-BN and population-BN. Evidence strength expressed as likelihood ratios: $EredHanded_{LR_+} = \infty$, $Ewallet_{LR_+} = 14.55$, $EseenCS_{LR_+} = 16.4$ (Evidence for being guilty).



AtCrimeScene T	AtCrimeScene F
$\frac{n}{N}$	$1 - \frac{n}{N}$

AtCrimeScene	EseenCS T	EseenCS F
F	0.035	0.965
T	0.951	0.049

AtCrimeScene	Guilty T	Guilty F
F	0	1
T	$\frac{1}{n}$	$1 - \frac{1}{n}$

Guilty	EredHanded T	EredHanded F
F	0	1
T	0.013	0.087

Guilty	Ewallet T	Ewallet F
F	0.020	0.980
T	0.291	0.709

Figure 2. Bayesian Network structure for using the opportunity prior with parameters n and N that are determined per run of simulation, opportunity-BN. Evidence strength expressed as likelihood ratios: $EredHanded_{LR_+} = \infty$, $Ewallet_{LR_+} = 14.55$, $EseenCS_{LR_+} = 27.2$ (Evidence for being at crime scene).

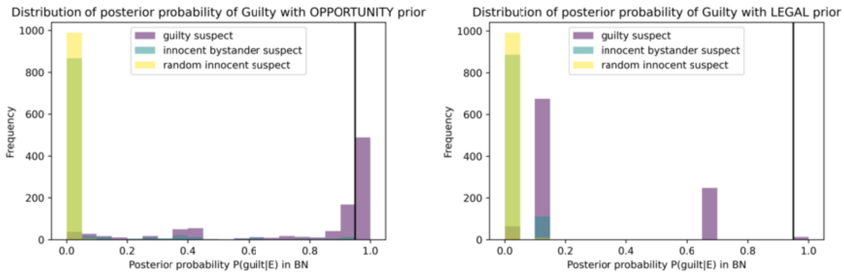


Figure 3. Frequency of posterior probabilities $\Pr(guilty|E)$ for (left) Opportunity-BN and (right) Legal-BN for three types of suspect: guilty, random innocent agent and innocent bystander, with a vertical line marking the threshold $t = 0.95$ of reasonable doubt.

3.2.1. Calculating the Posterior Probability of Guilt Given the Evidence

For every suspect in each run, we had three Bayesian Networks; one for each method of setting the prior and a set of evidence $e_{r,s}$, where e is the set of evidence found in run $r \in \{1,2,...,1000\}$ for suspect $s \in \{guilty, bystander, random\}$. For every suspect in each run, we first fine-tuned the opportunity-BN by setting n and N to the values found for that suspect in the run. The legal-BN and the population-BN were not changed. The evidence $e_{r,s}$ was entered into each BN to calculate the posterior probability $\Pr(guilty|e_{r,s})$. Calculations were performed using [14].

4. Results

For every suspect (s) in each run (r), posterior probabilities for the legal-BN (leg) $\Pr(guilty|e_{r,s})_{leg}$, the population-BN ($pop$) $\Pr(guilty|e_{r,s})_{pop}$ and the opportunity-BN (opp) $\Pr(guilty|e_{r,s})_{opp}$ are calculated. This results in 1000 posterior probabilities calculated per suspect. Figure 3 (left) shows a histogram of the frequency distribution of posterior probabilities as calculated by the opportunity-BN (Figure 2) for each of the three suspects. In all 1000 runs, the random innocent suspect has a posterior probability of guilt that is at most 0.05. The posterior of the innocent bystander ranges between 0 and 1, but is most frequently less than 0.05, and in two runs, this posterior is at least 0.95. Hence, both innocent suspects mostly have low posterior probabilities of guilt. For the guilty suspect, however, the posterior probability of guilt is greater than 0.95 in 489 out of 1000 runs. Hence, in nearly half of the runs, the $\Pr(guilty|e_{r,s}) > 0.95$ for the guilty suspect.

Figure 3 (right) shows the histogram of the posterior probabilities for the legal-BN (Figure 1). Compared to the BN with the opportunity prior, the variation in the posteriors is only due to the different evidence sets, and not a varying prior (which is always $\frac{1}{100}$, compared to the opportunity BN where the values of n and N, and hence the prior, as well as the evidence set, can vary per run). Both innocent suspects never have a posterior probability greater than 0.2. $\Pr(guilty|e_{r,s}) > 0.95$ for 13 out of 1000 runs. The results of the population-BN are not shown; they are similar to the results of the legal-BN.

We can consider the effect of each possible evidence set on the posterior probability for the opportunity-BN and the legal-BN for the guilty suspect (Figure 4) and for the innocent bystander (Figure 5). The boxplots for the opportunity-BN display a greater variation of values than the boxplots for the legal-BN; this is because, even though the weight of each evidence set is the same in both networks, the prior probability of guilt does not vary in the legal-BN, but varies in the opportunity-BN due to the different values of n (agents at CS) and N (agents with similar opportunity) in each run. When we look at the posterior probabilities per evidence set, we see that in evidence sets that contain $EredHanded = T$, for both the legal and the opportunity-BN, the posterior probability of guilt is always 1. This evidence only occurs for the guilty suspect and never for the innocent suspect.

We consider a set of evidence $Ewallet = T, Eredhanded = F, EseenCS = T$. This set of evidence can occur both for the guilty suspect as well as for innocent suspects. We see that for this set, in the opportunity-BN, the posterior probability $\Pr(guilty|Ewallet = T, Eredhanded = F, EseenCS = T) > 0.95$ is usually true for both the guilty suspect and the innocent bystander, as the box of the boxplot is past the 0.95 line. On the other hand, in the legal-BN, $\Pr(guilty|Ewallet = T, Eredhanded = F, EseenCS = T) = 0.7$. Hence, this is a combination of evidence for which the innocent bystander has a posterior probability of guilt beyond a reasonable doubt in the opportunity-BN, yet has a posterior probability of guilt of 0.7 in the legal BN.

To show the effect of a choice of threshold t for reasonable doubt on the performance of the Bayesian Networks as a classifier, a receiver-operating characteristic (ROC) curve was plotted. The ROC curve plots the true positive rate (y-axis) against the false positive rate (x-axis) for varying values of t. A true positive case is a guilty suspect with $\Pr(guilty|e) > t$ and hence is correctly classified as guilty. A false positive case is when $\Pr(guilty|e) > t$ for an innocent suspect, such that they would be incorrectly classified as guilty. A good classifier has a high number of true positives and a low number of false positives, which in a ROC plot is represented as a data point in the upper-left corner. We plot ROC curves for both opportunity-BN and legal-BN (Figure 6), where we take the false positive rate to be (red) the false positive rate for the innocent bystander, and (blue) the false positive rate for the random innocent. We test thresholds $0 \leq t \leq 1$ with steps of

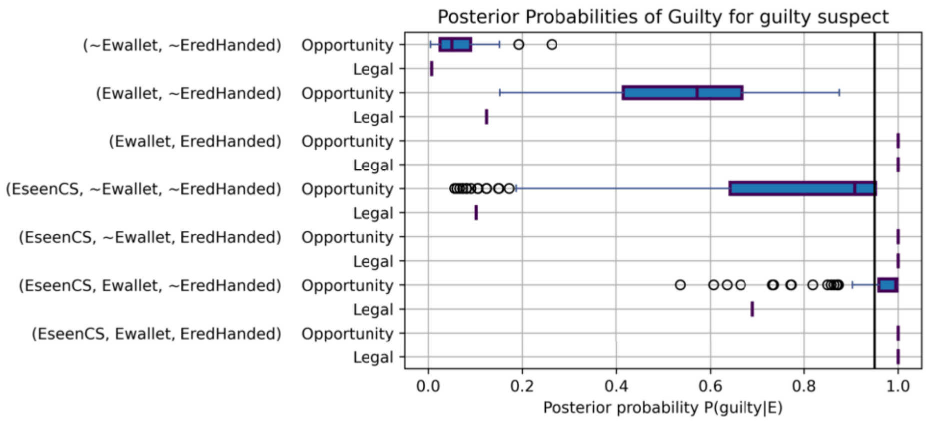


Figure 4. Boxplot showing, for every possible combination of evidence in the ABM, the posterior probability of guilt given that evidence, for the guilty agent, for opportunity-BN and legal-BN. The vertical line marks the threshold $t = 0.95$ of reasonable doubt.

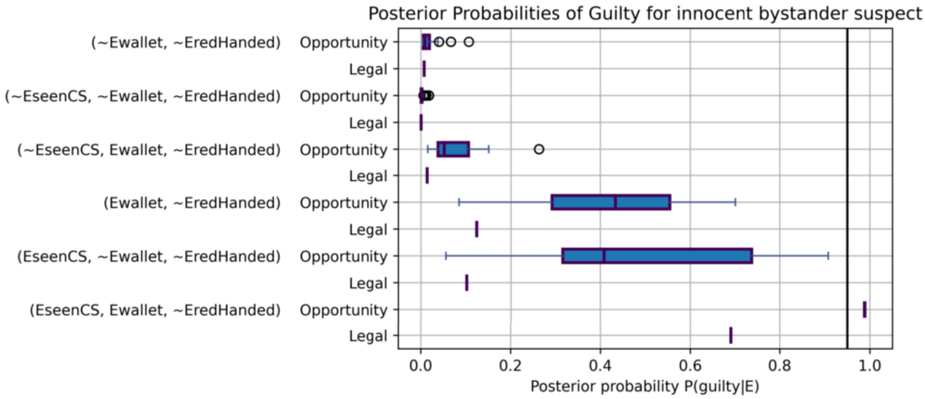


Figure 5. Boxplot showing posterior values per evidence set for the innocent bystander suspect for opportunity-BN and legal-BN. The vertical line marks the threshold $t = 0.95$ of reasonable doubt.

0.01. Points on the ROC curve marked with stars designate points at which $0.85 \leq t \leq 1$, all of which are relatively high thresholds. We see that in these cases, the false positive rate is near 0. For the legal prior, the true positive rate is also 0. For the opportunity prior, the true positive rate ranges between 0.2 and 0.7 for the innocent random agent at a false positive rate of 0. For the innocent bystander with the opportunity prior, we see a true positive rate of 0.7, yet this comes at the cost of a higher false positive rate than for the innocent random agent. Hence, the opportunity prior results in false positives: It incorrectly classifies innocent agents as being guilty. We see in general that the BNs performs better as a classifier for the innocent random agent than the innocent bystander. In general, we see that the best classification occurs at a threshold $t < 0.85$.

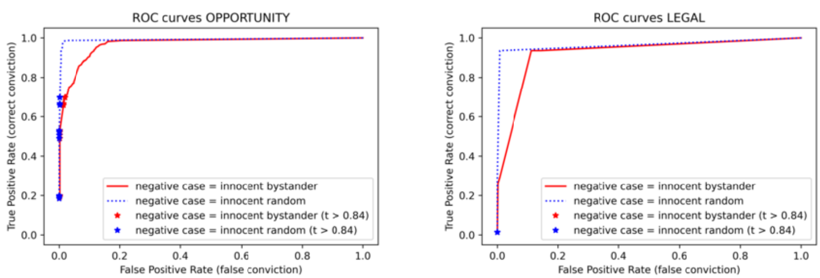


Figure 6. Receiver-operating characteristic (ROC) curves for both comparisons for both opportunity BN and legal BN. The positive case is the guilty agent, while the negative case is the innocent bystander (red) or the innocent random agent (blue). Stars (overlapping) designate points at which $0.85 \leq t \leq 1$.

5. Discussion

With a threshold of reasonable doubt set at $t = 0.95$, there is a difference in rates of true positives and false positives between the BNs that use the opportunity prior and the BNs that use the population prior and legal prior. We find that the opportunity-BNs better separate guilty from innocent suspects, yet, under some circumstances, would result in false positives for the innocent bystander under inconclusive evidence. Both the legal-BN and the population-BN have a low true positive rate and a low false positive rate, which means that they are ineffective, yet do no harm through false conviction. In the sequel, we consider the risk of false convictions, the background assumptions necessary for each method, and reasonable doubt.

What are the background assumptions for each method? Both the population prior and the opportunity prior assume that a crime has happened. The population prior assumes that a crime has happened within a certain territory (over which a population can be defined) and sets as prior the total number of agents in that total territory, which is in our case $N = 200$. The opportunity prior needs the most assumptions: In our implementation it requires specifically a precise location of the crime scene at a known time and to calculate N, we need to have one suspect, a certain identification of that suspect at a certain location and time, and know the number of N agents who were in the area A. Assuming that a crime has happened and that only one suspect was guilty leaves no space for integrating alternative scenarios. Additionally, a suspect should not be appointed without already establishing evidence, yet the process of picking exactly one suspect is not included in the model. The process through which these factors are established is left out of the formalism of the opportunity prior and it is unclear how reasoners and modellers can explicitly model sources of uncertainties about any of these factors. The legal prior makes no such assumptions. How these assumptions should be modelled is a question for future research.

When does the opportunity prior method risk false convictions? In [4], it was predicted that low values of n and N might give rise to false convictions and here we find that this is the case. Given the strength of available evidence in the simulation, we find that for all suspects that are convicted, innocent or guilty, it is the case that $n \leq 2$ and $N < 25$. However, there are also suspects with $n \leq 2$ and $N < 25$ that are acquitted when we do not find strong enough evidence for conviction (such as when $Ewallet = F, EredHanded = F$). Hence, low values of n and N of the opportunity prior do not necessarily result in false convictions, but might result in false convictions when the evidence set found is convincing towards the side of the prosecution, such as in the case of the innocent bystander who both has the wallet and was seen on camera near the crime scene. In the real world, where we are estimating n and N, under-estimating N and n will have a higher risk of false convictions than over-estimating them.

What should the threshold be? A choice of threshold implies a choice in trade-offs. If the threshold is set too high, then both innocent suspects and guilty suspects will be classified by the BN as innocent. If the threshold is too low, then both innocent and guilty suspects will be classified as guilty. There is always a trade-off between the number of true and false positives, however, what the balance between true and false positives should be is an open question. The ROC curves (Figure 6) show that, in case of this particular simulation, we can increase the number of true positives while not increasing the number of false positives equivalently. However, this point occurs at a threshold $t < 0.85$, which would be too low to represent ‘beyond reasonable doubt’. The question remains: Should the threshold of reasonable doubt be set to represent some optimal balance between convictions and acquittals, or because it represents some standard of evidence?

6. Conclusion

We created a statistical ground truth in an ABM to model a simulated theft. This ground truth is modelled in three types of Bayesian Networks, each using a different prior probability of guilt. We look at the posterior probability of guilt $\Pr(guilty|e)$ for innocent and guilty agents for every run. At a threshold $t > 0.95$, the opportunity-BN has a higher true positive rate than the legal- and population-BNs, however, it also has a non-zero false positive rate, in contrast to the other two types. How the rates of true positives and false positives should trade off is an open question. Additionally, the opportunity-BN requires more assumptions than the other two: We need to know that a crime has been committed, as well as requiring a known crime scene, time and suspect.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

[1] R. Urbaniak and M. Di Bello, "Legal Probabilism," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Fall 2021 ed., 2021.

[2] D. V. Lindley, "Probability and the law," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 26, no. 3, pp. 203–220, 1977.

[3] N. Fenton, D. Lagnado, C. Dahlman, and M. Neil, "The opportunity prior: a proof-based prior for criminal cases," *Law, Probability and Risk*, vol. 18, no. 4, pp. 237–253, 2019.

[4] C. Dahlman, "Determining the base rate for guilt," *Law, Probability and Risk*, vol. 17, no. 1, pp. 15–28, 2018.

[5] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[6] C. Vlek, H. Prakken, S. Renooij, and B. Verheij, "A method for explaining Bayesian networks for legal evidence with scenarios," *Artificial Intelligence and Law*, vol. 24, no. 3, pp. 285–324, 2016.

[7] C. Vlek, H. Prakken, S. Renooij, and B. Verheij, "Building Bayesian Networks for legal evidence with narratives: a case study evaluation," *Artificial intelligence and law*, vol. 22, no. 4, pp. 375–421, 2014.

[8] L. van Leeuwen, B. Verheij, R. Verbrugge, and S. Renooij, "Using agent-based simulations to evaluate Bayesian Networks for criminal scenarios," in *ICAIL - International Conference on Artificial Intelligence and Law, June 19-23, 2023, Braga, Portugal*, ACM, New York, NY, USA, 2023.

[9] D. A. Lagnado, N. E. Fenton, and M. D. Neil, "Legal idioms: a framework for evidential reasoning," *Argument and Computation*, vol. 4, no. 1, pp. 1–18, 2012.

[10] N. Fenton, M. Neil, and D. A. Lagnado, "A general structure for legal arguments about evidence using Bayesian Networks," *Cognitive Science*, vol. 37, pp. 61–102, Oct 2012.

[11] N. Fenton, M. Neil, B. Yet, and D. Lagnado, "Analyzing the Simonshaven case using Bayesian Networks," *Topics in Cognitive Science*, vol. 12, pp. 1092–1114, Mar 2019.

[12] J. Kadane and D. Schum, *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. John Wiley and Sons, 1996.

[13] U. Wilensky, "Netlogo," http://ccl.northwestern.edu/netlogo/, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999.

[14] G. Ducamp, C. Gonzales, and P. Wuillemin, "aGrUM/pyAgrum : A toolbox to build models and algorithms for probabilistic graphical models in Python," in *10th International Conference on Probabilistic Graphical Models*, vol. 138 of *Proceedings of Machine Learning Research*, (Skørping, Denmark), pp. 609–612, Sept. 2020.