Investigating the value of qualitative Bayesian networks of complete cases as "double-check" tools on traditional judicial reasoning: An exploratory study

Leya Hampson[1,*], Ludi van Leeuwen[2]

[1]*Department of Transboundary Legal Studies, Faculty of Law, University of Groningen, The Netherlands*
[2]*Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands*

Abstract

The overarching aim of this study is to empirically examine the feasibility and value of complete-case Bayesian network (BN) modelling within judicial deliberation. Building on critiques concerning the subjectivity and perceived overprecision of quantitative BN approaches, it explores whether qualitative BNs can serve as "double-check" tools on traditional judicial reasoning. Employing a sequential, mixed-methods design, comprising independent modelling, structured reflection, and collaborative comparison, two independent modellers constructed both qualitative and quantitative BNs of the same Dutch appellate verdict. The findings show that qualitative models can capture the essential reasoning structure of the court and assist in identifying implicit assumptions, incomplete dependencies and sources of uncertainty. Quantification significantly impacted the structure of the networks and highlighted the importance of precise and stable variable definitions to enhance transparency and interpretability. While they did not expose probabilistic fallacies, qualitative BNs aligned with the court's reasoning and revealed interpretive gaps, highlighting their heuristic rather than substitutive value.

Keywords

Bayesian networks, Legal Reasoning, Belief Updating

1. Introduction

In the Netherlands, an ongoing debate concerns both the feasibility and utility of constructing quantitative Bayesian Networks (BNs) of complete criminal cases. This debate has been fuelled by three instances in which such models were presented during trial; twice by the prosecution and once by the defence. In all three cases [1], the courts decided not to use their analyses (see [1] for an overview), citing concerns about the reliability of the Bayesian method [2].

A *Bayesian model of a complete criminal case* can take various forms, depending on several factors: the employed modelling method (e.g., Bayesian Networks or linear Bayes [2]), the level of detail included, and whether the approach is qualitative [3] or quantitative. It further depends on the individual(s) constructing the model (e.g., investigators, experts, prosecution/defence, or the court itself) and in which phase of the legal process the model is developed (e.g., during investigation, at trial, in the deliberation chamber, or as an element of the written verdict). Finally, the purpose for which the model is constructed may differ: from guiding investigations, structuring argumentation during trial, to assisting judicial reasoning or serving as a means of 'double-checking' a verdict.

In the present study, the focus lies on models constructed post-verdict by experts, serving as this latter "double-check" function. A complete case in this context refers to a model incorporating all

[1]The Hague Court of Appeals, 14 October 2015, ECLI:NL:GHDHA:2015:2859; Hertogenbosch Court of Appeals, 22 November 2016, ECLI:NL:GHSHE:2016:5165; Zeeland-West-Brabant Court, 24 May 2016, ECLI:NL:RBZWB:2016,360

[2]Bayesian network modelling is the standard method used, see Fenton et al. (2016) [3]; linear Bayes is used by Alkemade in his case analyses of ECLI:NL:GHDHA:2015:2859 and ECLI:NL:GHSHE:2016:5165

[3]Qualitative Bayesian Networks, in this work, refer to the qualitative aspects of a BN: graph and variables, without a distribution, not to Qualitative Probabilistic Networks.

items of evidence explicitly mentioned by the court in its verdict. Any fact or piece of information cited by the court as part of its reasoning, irrespective of (un-)assigned probative value, is considered part of the evidential set to be modelled. The inclusion of such information signals it played a role in the court's assessment and should therefore be included in the model. Model completeness also involves addressing the relationships between various pieces of evidence, thus all dependencies should be explicitly modelled.

Several recurring arguments have been raised against the use of BNs in such comprehensive applications. Constructing a BN of a criminal case requires both statistical and domain-specific expertise, including an understanding of forensic evidence, investigative processes, and legal reasoning [2]. Furthermore, at present, no uniform method exists for constructing BNs of criminal cases [4, 5, 2, 6, 7], and thus modellers may differ in how they define variables, determine dependencies, and structure the overall narrative of the network (e.g., employing an idiom-based approach [5] vs a scenario-based approach [4]). These variations introduce an element of subjectivity [8], but these concerns become especially pronounced at the quantification stage. Critics question the basis on which numerical probabilities are assigned to the conditional probability tables – often summarised as the question *where do the numbers come from?* [9]. In practice, the assignment of prior and conditional probabilities relies heavily on expert judgement or subjective estimation rather than on empirical data [10, 11]. Even when empirical data is available for constructing Bayesian networks, some variables inevitably concern unique events – such as the probability that a particular defendant performed a specific act – for which no empirical frequencies can exist [8, 12]. Critics argue that quantification can therefore convey a misleading impression of precision, concealing subjective assumptions behind a facade of mathematical rigour [6, 13]. This perceived precision can, in turn, lend the model undue persuasive force [3]. Judges, typically untrained in Bayesian principles, may overvalue numerical outcomes or rely too heavily on expert interpretation. Such reliance, whilst often unavoidable, can blur the boundary between evidential assessment and external expertise, potentially undermining judicial independence [14]. Moreover, the modeller's interpretive framework may subtly shape how the court perceives and weighs the evidence.

These critiques have led some commentators to question the role that Bayesian reasoning can and should assume in legal decision-making [3]. Rather than serving as a comprehensive analytical framework, it is argued that Bayes should be employed in the qualitative and global sense [9, 15]; as a means of double-checking or triangulating existing legal reasoning. In this view, the function of Bayesian modelling is not to determine the quantitative outcome of a case, but to ensure that no probabilistic inconsistencies or fallacies have occurred during the deliberation process.

While much of the existing discourse on Bayesian modelling in law has focused on its limitations, these critiques can overshadow the potential value of the approach. In particular, debates surrounding subjectivity, numerical precision, and the source of probabilities have come to dominate the discourse, leaving comparatively little attention to the qualitative insights that the modelling process itself can offer (see Meester (2020) [6], for instance, who briefly touched upon the qualitative value of the proposed BN before delivering a 22 page analysis on why the numbers did not work). By focusing narrowly on the reliability of numerical outputs, critics risk overlooking the interpretive and diagnostic functions that Bayesian reasoning may serve in structuring legal argumentation and revealing inconsistencies in evidential assessment.

The current study forms part of a broader project that aims to empirically examine the feasibility and value of complete-case Bayesian modelling within judicial deliberation. It builds directly on the theoretical claim that, given the critiques surrounding subjectivity and feasibility, Bayes may have a more limited—but still valuable—role to play in legal reasoning: not as a quantitative decision-making tool, but as a qualitative means of checking the coherence of evidential reasoning.

The study therefore explores whether qualitative Bayesian networks can, in practice, fulfill this proposed "double-check" function while sidestepping the main critique in the literature: the subjectivity inherent in quantification. Using a Dutch appellate case as a test example, two independent modellers constructed both qualitative and quantitative Bayesian Networks (BNs) of the same verdict. This design allows us to examine how quantification influences the structure and interpretation of the networks, and whether the qualitative models alone capture the essential reasoning structure necessary for judicial

evaluation.

Guided by this objective, the present paper makes an initial step toward addressing the following three exploratory research questions:

1. Do the modellers perceive their models as complete representations of the case?
2. To what extent does quantification impact the structure of the network?
3. Can a qualitative network effectively serve as a "double-check" tool on traditional legal reasoning? More specifically, we examine the ability of qualitative networks to a) detect errors, b) align with the courts reasoning, and c) provide additional value beyond identifying probabilistic fallacies.

2. Methods

2.1. Design

The study employed a sequential, mixed-methods design consisting of (i) independent modelling, (ii) structured reflection, and (iii) collaborative comparison. The two modellers were the authors themselves, both PhD candidates with prior experience in Bayesian modelling (for more detailed information on their respective backgrounds and relevant modelling experience please see Appendix A).

2.2. Case material

To preserve the blindness of the modellers to the case pre-modelling, an independent external expert selected the case and translated the appellate verdict into English to facilitate both modellers understanding of the case (the full translated verdict document is available as supplementary material and will be made available upon request). The case [4] concerns the armed robbery of a supermarket in Nieuw-Dordrecht in March 2021, for which the defendant was convicted. The evidential material described in the verdict includes CCTV footage, eyewitness testimonies, gait observations, weapon evidence, and cell-site data. No individual item carried strong probative value; instead, the conviction relied on the cumulative effect of multiple weaker indicators. This made the case suitable for examining how BN modelling may support or "double-check" judicial reasoning in borderline or uncertainty-sensitive cases. The modellers were restricted to the evidence explicitly mentioned in the court's verdict. This reflects judicial practice, in which the judge has the exclusive authority to select and weigh evidence.

2.3. Procedure

The full modelling process consisted of eight structured phases [5], illustrated in Figure 1. Only data collected during the first six phases are discussed in the current analysis.
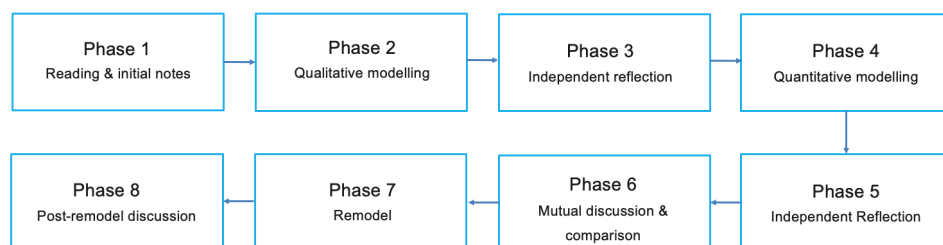

Figure 1: Overview of modelling phases and study procedure

---

[4]Arnhem-Leeuwarden Court of Appeal, 15 May 2025, ECLI:NL:GHARL:2025:3013.
[5]The modelling process was intentionally divided into distinct, sequential steps. While this may not reflect real-world modelling practices, in which modellers may iterate between qualitative structure and quantitative input, or develop both simultaneously, it allowed for a more controlled evaluation of the impact of quantification.

Phase 1 required modellers to independently read the case and make initial notes. In phase 2, they individually constructed a qualitative BN of the case, before, phase 3, completing a structured reflection on the process of constructing the qualitative model. In phase 4, each modeller expanded their qualitative structure into a quantitative network, and in phase 5 completed a second structured reflection. In phase 6, the modellers collaboratively compared and discussed their final BNs using a structured discussion protocol. Phases 7 and 8 involved the joint construction and discussion of a single, agreed-upon BN.

The semi-structured reflection and discussion protocols (available in the supplementary material), employed in phases 3,5,6 and 8, were developed to introduce a degree of procedural consistency into this exploratory study, providing a systematic framework for within-subject (and later between-subject) comparison.

Throughout all phases, the modellers followed a think-aloud-protocol to document the reasoning behind their modelling choices. All sessions were recorded via Zoom Workplace for Education (Version 6.5.12), which provided video, audio and automatic transcription. The time spent on each step was systematically registered. All models were constructed using the Hugin software package (Version 9.5 & 9.6) [6].

Although both modellers intended to use the same version of Hugin, one inadvertently used the Pro edition while the other used the free version. This did not have a significant impact on the modelling process or the planned analysis; all functionalities and comparison measures required for the study are available in both editions. The main difference concerns the resulting model complexity, which does not affect the outcomes reported here.

3. Results

This section presents the constructed models, contributing to the assessment of model completeness. We compare the structural changes between the qualitative and quantitative networks, evaluating the value of qualitative case modelling compared to its quantitative counterpart.

3.1. Model overviews

While both models were fully quantified, only their structural components are reported here, as the analysis focuses on the impact of the quantification process itself rather than on the specific numerical probabilities assigned. The completed conditional probability tables are available upon request and will be addressed in forthcoming publications.

3.1.1. Modeller 1

Modeller 1 employed an idiom-based modelling approach, as outlined in [5]. The qualitative and quantitative models can be seen in Figure 2a and 2b respectively [7].

Both networks are centred around the ultimate hypothesis concerning the defendant's guilt (*H_Defendant_Guilty*). Supporting evidence is organised around four sub-hypotheses —relating to motive (*Hm_Defendant_Had_Motive*), opportunity (*Ho_Defendant_had_opportunity, H_D_at_crime_scene*), gun evidence (*H1_Defendent_gun_used_in_robbery, H1_D_gun_used_in_robbery*), and car evidence (*H2_RobberyCar_belongs_defendant, H_D_man_in_CCTV*) — each serving as a parent to the ultimate hypothesis. Two items of evidence, eyewitness testimonies, are directly connected to the ultimate hypothesis.

The total time spent modelling was 188 minutes (3.1 hours); 82 minutes (1.4 hours) were spent on reading the case and making initial notes, 54 minutes (0.9 hours) on building the qualitative structure and 52 minutes (0.9 hours) on the quantification process.

---

[6]Hugin was selected due to its ease of use and accessibility, making it a suitable platform for potential employment in real-world legal settings.

[7]The models presented in this section are reduced in size for readability, please see Appendix B for the full scale versions.

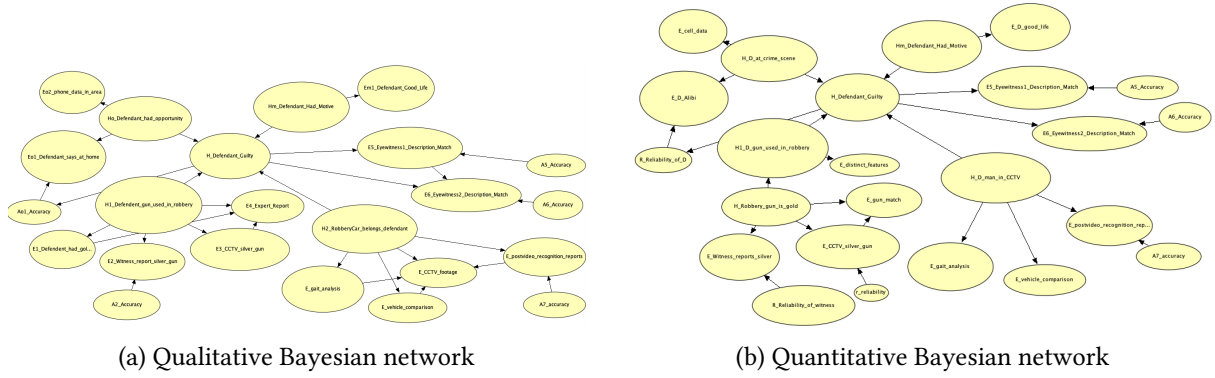(a) Qualitative Bayesian network   (b) Quantitative Bayesian network

Figure 2: Modeller 1: comparison of qualitative and quantitative Bayesian networks.

During the quantification process, several structural changes were made to the original qualitative network. In the quantified network, an additional sub-hypothesis relating to the gun evidence was introduced (*H_Robbery_gun_is_gold*). The total number of evidence nodes decreased by one, while the number of reliability nodes increased by one. A side-by-side comparison of the basic network properties can be seen in Table 1.

Table 1
Network statistics for both modellers

Property	Modeller 1		Modeller 2	
	Qualitative	Quantitative	Qualitative	Quantitative
Total number of nodes	23	24	52	52
Total number of edges	29	25	57	56
Total number of hypotheses	5	6	16	19
Total number of evidence nodes	13	12	28	26
Total number of reliability nodes	5	6	8	7

3.1.2. Modeller 2

Modeller 2 adopted a temporal-narrative (scenario-based) modelling approach, based on [4]. The qualitative and quantitative networks are presented in Figure 3 and 4 respectively.

The qualitative model features two parallel structures, separating hypotheses specific to the known defendant (red nodes), abstract hypotheses specifying the criminal events, without reference to a specific defendant (yellow), and a set of identification nodes (blue) linking these two layers. The ultimate hypothesis is *DefendantRobsShopThreatensetc*, the subhypotheses of the scenario are temporally ordered around this node. The evidence (green) represented explicitly allows for the possibility that the crime happened, without the defendant being guilty. Reliability and alternative explanation considerations are represented in orange.

The total modelling time was 9.5 hours; 2.5 hours were spent on reading the case, 3 hours on building the qualitative structure and 4 hours on the quantification process.

In the quantitative model, the identification nodes were streamlined, and several links within the abstract criminal scenario were removed. Two evidence nodes related to the tattoo cluster were removed. Additionally, the structure of the gun evidence cluster underwent substantial modification. A simple analysis of structural changes between the qualitative and quantified models reveals no difference in the number of nodes. The number of edges decreased slightly, from 57 to 56. Similarly to modeller 1, the number of hypotheses decreased between models, and the number of evidence nodes increased. A full list of network properties can be seen above in Table 1.
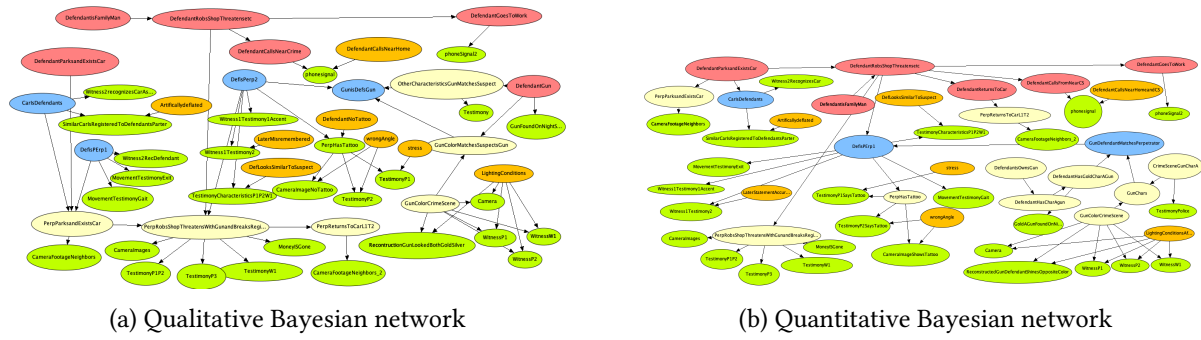
(a) Qualitative Bayesian network   (b) Quantitative Bayesian network

Figure 3: Modeller 2: comparison of qualitative and quantitative Bayesian networks.

3.2. Model revisions during quantification

To elucidate the model alterations introduced during the quantification process, we analyzed the differences in node composition and link configuration between the two networks. This qualitative analysis serves to capture changes that are not apparent from numerical comparison alone (e.g., models may exhibit identical node counts while representing distinct variable sets). Nodes were classified as *added* or *removed* when introduced or omitted between versions. A variable name was considered *refined* only when one or more words were changed (minor adjustments, such as abbreviations, renumbering or variations in capitalisation were disregarded). A *definition change* was coded when a node's conceptual scope shifted (e.g., when a variable was broadened, narrowed or reframed within the same evidential theme; this includes negation). Alterations to nodes, whether additions or removals, inherently necessitate corresponding modifications to the network's edges. Link changes were accordingly classified as *added*, *removed*, *direction change* or *mediated* (i.e. when an existing connection became indirect through the introduction of an intermediate node).

3.2.1. Modeller 1

Of the 23 variables in the initial qualitative model, 19 remained in the quantitative expansion. Three variables were added: one hypothesis (*H_Robbery_gun_is_gold*), one piece of evidence (*E_distinct_features*), and a reliability consideration (*R_reliability*). Two evidence nodes, *E_CCTV_footage* and *E1_Defendant_had_gold_gun*, were removed. Six node labels were refined, and one variable was redefined to reflect a conceptual shift in meaning. A full qualitative overview of network modifications is presented in Appendix A, a summarised version presenting the frequency count of each modification can be seen in Table 2.

Only one link was removed independently of node changes: the connection between the two eyewitness testimony nodes, *E5_Eyewitness1_description_match* and *E6_Eyewitness2_description_match*. The removal of one evidence item, *E_CCTV_footage*, directly led to the deletion of its four links to parent nodes. Likewise, the removal of a further evidence node, *E1_Defendant_had_gold_gun*, resulted in the automatic loss of two additional links. The inclusion of the sub-hypothesis *H_robbery_gun_is_gold* not only necessitated the inclusion of a link to the existing network but also reconfigured several existing connections. Relationships between variables that were previously direct — such as those between *H1_Defendant_gun_used_in_robbery* and multiple evidence nodes—became mediated through this intermediate node. Table 2 outlines these link changes, distinguishing between those arising as a direct consequence of node alterations and those resulting from independent structural revisions.

3.2.2. Modeller 2

There are 34 variables that are the same across both models, and there are 18 variables in the qualitative model that are not in the quantitative, and vice versa. The differences in node composition can be summarised as follows: one identification node was removed to streamline the model; eight scenario

Table 2
Modeller 1: Summary of variable and edge modifications between qualitative and quantitative networks.

(a) Variable modifications by node type				(b) Edge modifications by relation to node change				
Node Type	Added	Removed	Name Refined	Definition Changed	Relation	Added	Removed	Mediated	Total
Hypothesis	1	0	1	1	No	0	1	0	1
Evidence	1	2	3	0	Yes (Child added)	1	0	0	1
Reliability	1	0	2	0	Yes (Child removed)	0	5	0	5
					Yes (Parent removed)	0	1	0	1
					Yes (Parent added)	2	0	0	2
					Yes (Sub-hypothesis added)	0	0	3	3
Total	**3**	**2**	**6**	**1**	**Overall**	**3**	**7**	**3**	**13**

(sub-hypothesis) nodes were refined; and two reliability nodes were modified to clarify their meaning. Within the tattoo evidence cluster, two evidence nodes were removed: the node *PerpHasTattoo* previously mislabelled as evidence was reclassified as a supporting hypothesis, and *DefendantNoTattoo* was deemed redundant, as this condition was already ensured by the CPTs. Several evidence labels were refined for greater specificity, most notably within the gun evidence cluster.

There were significant changes in the arcs of the network. For a full overview, see Appendix A. The main changes are in the abstract scenario structure. This included edges between these abstract scenario nodes. This was deemed unnecessary in the quantitative model, as these relations were captured in the specified scenario.

Table 3
Modeller 2: Summary of variable and edge modifications between qualitative and quantitative networks.

(a) Variable modifications by node type				(b) Edge modifications by relation to node change				
Node Type	Added	Removed	Name Refined	Definition Changed	Relation	Added	Removed	Dir. Changed	Total
Scenario	3	0	5	1	No	2	7	0	9
Evidence	0	2	7	1	Yes (Child added)	3	0	0	3
Reliability	0	0	1	1	Yes (Parent added)	3	0	0	3
Hypothesis	0	1	0	0	Yes (Re-specified)	0	0	1	1
Total	**3**	**3**	**13**	**3**	**Overall**	**8**	**7**	**1**	**16**

3.3. Discussion protocol analysis

Table 4 summarises the modellers' reflections regarding the alignment between their Bayesian networks (BNs) and the court's reasoning. Despite being instructed to model the complete case, neither modeller felt that their BN fully represented all evidence cited in the verdict. Both, however, expressed confidence that their networks accurately captured the key inferential relationships and mirrored the overall logical structure of the court's reasoning. Both modellers reported that the (in)dependencies within their respective networks were clearly defined, though Modeller 2 qualified this assessment by noting that full clarity would only be achievable after quantification of the conditional probability tables (CPTs) (for the complete reflective responses see the Supplementary Materials [8].

Similarly, both modellers confirmed that alternative hypotheses, including negations and competing scenarios, had been considered conceptually, though only Modeller 2 incorporated these alternatives explicitly into their network structure.

To further explore the potential of qualitative models to serve as a "double-check" on traditional judicial reasoning, the modellers were asked to reflect on the error detection capabilities of their networks. The summarised responses are presented in in Table 5. Neither modeller was able to explicitly identify probabilistic or logical fallacies within the court's reasoning. Modeller 1 clearly stated that the network did not implicitly expose any probabilistic fallacies, whereas modeller 2 emphasised that such

---

[8]Available upon request.

Table 4

Alignment between Modeller 1 and Modeller 2 Bayesian Networks

Alignment Question	Modeller 1	Modeller 2
Does the BN fully represent all evidence cited in the court's verdict?	No	No
Are all (in)dependencies clearly defined?	Yes	Yes
Is the network sufficiently nuanced for the analysis it is meant to support? Are all chains of inference visible?	No	No
Have alternative hypotheses been considered, including both the negation of the main hypothesis and any competing hypotheses?	Yes	Yes
Does the structure mirror the logic of the court?	Yes	Yes

assessments were not possible at the qualitative stage of modelling. Both agreed that the direction of inference within their networks was plausible and consistent with real-world causal relationships.

Table 5

Error Detection Reflections for Modeller 1 and Modeller 2

Error Detection Question	Modeller 1	Modeller 2
Does the BN implicitly expose any probabilistic fallacies? If so, which ones?	No	Unsure
Are any logical fallacies or other reasoning errors (beyond calculation mistakes) present? If so, which one(s)?	No	Unsure
Is the direction of inference plausible, i.e., are the directions of the links between the nodes consistent with the real-world causal relationships they represent?	Yes	Yes
Were any plausible alternative hypotheses ignored in the court's verdict?	No	Unsure
Does building the BN expose ignored alternative hypotheses?	No	Yes

Table 6 summarises the modellers' reflections on the broader value of constructing a qualitative BN, beyond its potential for explicit error detection. Both reported that the process deepened their understanding of the case, helping them visualise dependencies, recognise uncertainty sources, and detect potential "jumps" in the court's reasoning.

Table 6

Modellers' Reflections on the Value of Constructing a Qualitative BN

Value considerations	Modeller 1 (full response)	Modeller 2 (full response)
How did building the qualitative structure increase your understanding of the case?	It forced me to think a lot deeper about the (possible) dependencies between the different pieces of evidence, greatly enhancing my understanding of the case. I would not have had such deep internal debates on dependencies without modelling.	Yes, although I was sometimes confused about how to model aspects such as "identification." The BN clarified the "sources" of uncertainty.
To what extent did Bayesian thinking assist the identification of missing evidence?	Bayesian thinking helped identify missing evidence. Additional CCTV footage could have strengthened links regarding the defendant's car; more information from the supermarket might have clarified uncertainties.	There appeared to be missing evidence for alternative explanations, issues with witness 2's reliability, and gaps in the narrative surrounding the defendant's car and gun.
To what extent did Bayesian thinking assist the handling of (in)dependencies?	Very much so. Bayesian thinking forces deeper consideration of (in)dependencies, and visualising these via links clarified relationships I had previously overlooked.	I think something is wrong with the gait/movement testimony regarding independence.
Did building the qualitative model help you expose any jumps in reasoning?	Yes. The match between the defendant's car and the CCTV car seems a jump; additional evidence is needed. Likewise, assumptions about the gun require frequency information to justify the match.	Yes. The verdict assumes various sightings relate to the defendant without justification. The alternative explanation for the phone mast location was not considered. There is also an unjustified assumption about the identity of the man in the store and the man near the car.

4. Discussion

4.1. On the feasibility of constructing qualitative models of complete cases

While this article does not aim to make generalizable claims about the feasibility of BN modelling in judicial deliberation, two key considerations emerged: model completeness and time constraints.

Modeller 1 was restricted by the Hugin free version, which allows for a maximum of 50 state nodes. As a result, subjective decisions had to be made regarding which evidential items to include in the network, resulting in an incomplete representation of the court's verdict. This limitation conflicts with the judicial principle that the judge has the exclusive authority to weigh and select the evidence, making the Hugin free version (Hugin Lite 9.6), although accessible, an unsuitable platform for complete case modelling. Model 2, by contrast, offers a more comprehensive representation of the case. The only notable omission concerns the car investigation, where not all witness testimonies were explicitly modelled. However, the reading of the case and construction of the qualitative structure alone required around 5.5 hours, which — following verbal discussions with legal professionals — may be considered as unfeasible in the context of double-checking the verdict in the deliberation chamber.

Post-hoc reflection further revealed that neither modeller felt that their networks fully captured all considerations of (in)dependencies: some were omitted due to uncertainty on how to operationalise them, while others were only recognised in later mutual discussion. The court's verdict provided both modellers with difficulty in this aspect: dependencies were rarely explicitly discussed, and missing information (potentially available to the court but not reflected in the verdict) would have been required to model such relationships accurately.

Both modellers' experience reflect longstanding discussions in the literature concerning the definition and attainability of *model completeness* in Bayesian modelling of legal cases, namely, that legal reasoning, as expressed in verdicts, is inherently selective and often omits the explicit dependencies required for full formal representation. The study therefore provides empirical support for these discussions by showing that the limits to completeness arise not only from technical or time constraints, but from the nature of judicial reasoning itself.

4.2. Understanding model modifications between networks

Our second research question aims to examine the extent to which quantification impacts the structure of a qualitative BN. The extensive structural revisions observed during the quantification process indicate that the process of assigning probabilities did far more than simply numerically parametrise a pre-defined network. It reshaped how both modellers conceptualized the case. While it may initially appear that such modifications arose simply from prolonged cognitive engagement with the case material, both modellers explicitly rejected this interpretation. They identified quantification itself – the act of populating conditional probability tables and directly confronting questions of the type *What is the probability of this event, given this evidence?* – as the moment that prompted then to reevaluate their earlier assumptions and adjust the structure accordingly.

Modeller 1's transcript (request supplementary material) illustrates this process. While completing the CPT for *H_Defendant_gun_used_in_robbery*, they paused mid-sentence:

> "So, if the robbery gun is gold ... Oh, I have just realised I would like to add nother node, actually, for the gun ..."

Here, the demand to specify probabilities directly triggered recognition of an unmodelled conceptual distinction, prompting structural refinement. A similar process occurred during the quantification of the eyewitness testimonies:

> "Ok, now, actually, I am not too sure about this dependent link right here between the two eyewitnesses. So I am going to remove it ..."

Modeller 2's transcript (supplementary material) demonstrates a similar dynamic. While filling in the CPTs for the car-related CCTV evidence, they repeatedly interrupted themselves to reconsider the dependencies at play:

> "This is ... the probability, given that the guy breaks the car and gets out of the car, what is the probability that we see that in that location?" ... Well, we think that is pretty likely ... Oh yeah, because this one is still weird, ... it has an extra parent..."

These reflections show that probability elicitation can act as a diagnostic probe into model coherence and completeness, highlighting uncertainties which remained hidden during qualitative construction. Across both modellers, despite the study design, quantification was not experienced as a separate technical phase but an integral part of understanding the case. When asked directly whether improved understanding resulted from quantification or simply from spending more time with the model, both answered unequivocally: *Quantification.* As modeller 1 reflected:

> "If I just continued thinking about the structure without the numbers, I would not have realised these things ... It was actually when I went to fill it in, and I was putting up these questions in my head, 'Ok, if he is a family man, how likely is it that...', and then I was like, wait, how am I meant to say that?"

Modeller 2 echoed this point:

> "But the only way you can get at this, like, qualitative view is through, ... is by going through the quantitative view.... I don't think, at least the way I approach it, I can separate the qualitative and the quantitative part."

While previous studies have discussed the broader epistemic value of BN modelling [3, 6], they do not distinguish between the qualitative and quantitative phases of the modelling process, nor assess the individual contribution of each. Our study makes this distinction explicit by comparing networks before and after quantification. Taken together, the above accounts show that quantification functioned as a source of conceptual change. This raises an important question for the function of qualitative BNs as post-hoc "double-check" tools on judicial reasoning: if the act of quantification substantially alters the structure and consequential interpretation of the model, can an unquantified network truly serve as a reliable check of the court's reasoning and hidden assumptions?

4.3. The importance of clear node definitions

An important insight emerged during the modelling and discussion phases: proper and consistent variable definitions are essential for transparency and interpretability in BN modelling of complete criminal cases. During quantification, modeller 1 refined six variable names (representing 26% of the total) and redefined one variable, while modeller 2 revised thirteen variable labels (25%) and redefined two. These are substantial proportions of the variable set. Thus both modellers not only encountered difficulties maintaining clear conceptual clarity across modelling stages, as well as when discussing their respective models, repeatedly requiring clarification of each other's variable meanings despite working from the same evidential basis (as reflected in the discussion transcript, available as supplementary material). This issue was explicitly acknowledged by the modellers in their post-quantification discussion:

> Modeller 2: "… I think that this is a big problem of Bayesian network [modelling] that … is not discussed, but, …, if you make a [node], and you fill out the table with one sort of interpretation of what that variable means, and then maybe later on you look at it and you forgot what exactly you [meant], and then you fill out the difference. Or … you go on … add another node, and … [forget] exactly what you [meant] by the first interpretation. Or … you condition on the parents, but maybe you now consider the parents as broader than before… there is … a sort of implicit inconsistency in how you define the nodes."

> Modeller 1: "I … mid-quantification … changed all my node names, because I said I need to be more specific. … I'd write … a general statement —- CCTV —- what does CCTV mean?"

This underscores a broader issue rarely reflected in the literature: while BNs make evidential structures explicit, their interpretability ultimately depends on precise, stable variable semantics. Fenton et al.

(2016) [3] emphasize that the strength of a BN lies in its capacity of represent complex evidential variables transparently; yet this transparency collapses if variables are ambiguously defined or evolve mid-modelling. Their later analysis of the Simonshaven case [16] provides a clear example: although the paper presents a complete BN and describes the modelling process in detail, it offers little explanation of how individual variables were defined. As a result, the model's structure can be difficult to interpret, even for technically informed readers. Should Bayesian modelling, whether qualitative or quantitative, serve as an independent, structured "double-check" on the court's traditional reasoning process, such clarity and definitional precision are essential.

4.4. The value of qualitative networks as "double-check" tools

It is difficult to empirically evaluate whether qualitative networks can serve as double-check tools. in an attempt to somewhat formalise this analysis, in our discussion checklist we evaluated the qualitative nets based on various criteria, amongst others: alignment, error detection and value.

Both modellers reported that constructing the qualitative network substantially enhanced their understanding of the case. The process encouraged explicit consideration of evidential dependencies and helped identify areas of uncertainty or missing information. As Modeller 1 noted:

> "It forced me to think a lot deeper about the (possible) dependencies between the different pieces of evidence"

Modeller 2 observed that the BN:

> "identifies the 'source' of uncertainty."

These reflections confirm the interpretive and diagnostic value of qualitative BNs: they make reasoning structures visible and expose where uncertainty resides. This finding differs from the existing literature in that we explicitly distinguish between the qualitative and quantitative value of the networks, rather than treating the BN modelling process as a single tool.

However, neither modeller identified explicit probabilistic fallacies or logical inconsistencies in the court's reasoning. This absence does not necessarily undermine the potential of BNs as "double-check" tools, as the case itself may have been "perfectly" reasoned, but, together with the observed structural impact of the quantification, highlights a potential limitation warranting further investigation: Qualitative modelling alone may be insufficient to detect more subtle probabilistic missteps.

Beyond explicit error detection, both modellers agreed that the qualitative process revealed implicit assumptions and unspoken leaps in the court's reasoning. Modeller 1 highlighted the treatment of the car and gun evidence—particularly the strength of the assumed matches and the absence of frequency information—as examples where the court appeared to rely on under-specified or weakly supported inferences. Modeller 2 drew attention to gaps in the treatment of identification evidence, questioning the implicit assumption that multiple different eyewitness descriptions referred to the same individual. These observations suggest that qualitative modelling can surface areas where reasoning relies on implicit or weakly articulated assumptions, even when no formal probabilistic fallacies are present. This directly aligns with Prakken's (2020) [2] argument that the use of Bayes in law should not dictate conclusions but rather support structured dialogue on evidential coherence. The qualitative BN accomplishes precisely that. It renders visible the tacit dependencies that traditional judicial writing leaves implicit, thereby allowing others to ask whether the inferential links assumed by the court are defensible, complete, and mutually consistent.

Together, the findings suggest that the primary value of constructing qualitative BNs lies not in error detection per se, but in fostering a structured form of reflection of evidential coherence. As "double-check" tools, their strength is heuristic rather than diagnostic: they do not mechanically verify the correctness of a verdict but create a structured environment in which the assumptions, dependencies, and uncertainties embedded in judicial reasoning can be examined explicitly. This reflective capacity is particularly relevant in appellate or review contexts, where transparency of reasoning is as important as its substantive outcome.

4.5. Limitations and future research

The present study involved only two expert modellers (the authors) and focussed on a single criminal case. Consequently, the insights derived from this work are exploratory in nature and cannot be generalised to the wider population of legal practitioners or to other case types. It forms part of a broader study examining the feasibility and value of constructing Bayesian networks of complete criminal cases. Future work will address several outstanding questions, including whether two independent modellers analysing the same case can arrive at the same (or similar) outcomes, and whether disagreements — if any — can be resolved through discussion and model refinement. Further, drawing on the insights and difficulties encountered in this study, the authors aim to develop and test more formalised BN comparison tools (including the development of the discussion protocols into a standardised BN evaluation tool, supported by a taxonomy of model modifications and their potential interpretive and structural implications). To extend these findings, future research should apply the experimental approach to additional criminal cases and a broader participant group with varying expertise—from students to legal practitioners and forensic advisors. Within the wider PMJ project, the long-term goal is to develop tools that help judges recognise and avoid probabilistic reasoning fallacies. If complete-case modelling proves impractical due to time or complexity, simplified alternatives become essential. Ongoing work explores a scenario-based method and a complementary question list designed to support structured probabilistic reflection.

5. Conclusion

This study provides an initial empirical examination of qualitative Bayesian networks as post-hoc "double-check" tools in legal reasoning. By contrasting qualitative and quantitative models of a single appellate case, we found that the quantification process significantly influenced not only network parametrization but also conceptual understanding of the case and structural formulation. Although qualitative networks alone can clarify evidential dependencies, expose implicit assumptions, and enhance transparency in reasoning, they thus far appear limited in their capacity to identify probabilistic fallacies or subtle logical errors without numerical specification. Our findings underscore the dual nature of Bayesian modelling in legal settings: its strength lies in structuring complex evidential relations, yet its interpretive reliability depends on clear, consistent node definitions and an awareness of how quantification reshapes conceptual framing. Based on the current case analysis, given the time and expertise required for the complete-case modelling, the practical use of qualitative BNs in judicial deliberation may lie in their role as heuristic or educational tools, supporting judges and legal practitioners in identifying uncertainty and hidden assumptions rather than in producing decisive probabilistic outcomes. Future work should extend this approach to additional cases and participants, standardize reflection and comparison protocols, and further explore hybrid frameworks that balance qualitative transparency with the analytical rigour of quantified reasoning.

Acknowledgments

6. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 5.0 in order to: Grammar and spelling check, paraphrase, and reword. After using this tool, the authors reviewed and edited the content as

needed and assume full responsibility for the content of the publication.

References

[1] A. R. Mackor, Bayesian modelling of criminal cases as a whole: A philosophical reflection on dutch case law, Questio Facti (forthcoming, 2026).

[2] H. Prakken, A new use case for argumentation support tools: supporting discussions of bayesian analyses of complex criminal cases., Artif Intell Law 28 (2020) 27–49. doi:10.1007/s10506-018-9235-z.

[3] M. N. Norman Fenton, D. Berger, Bayes and the law, Annu Rev Stat Appl (2016) 51–77. doi:10.1146/annurev-statistics-041715-033428.

[4] C. Vlek, H. Prakken, S. Renooij, B. Verheij, Modeling crime scenarios in a Bayesian network, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, ACM, Rome Italy, 2013, pp. 150–159. URL: https://dl.acm.org/doi/10.1145/2514601.2514618. doi:10.1145/2514601.2514618.

[5] D. A. Lagnado, N. Fenton, M. Neil, Legal idioms: a framework for evidential reasoning, Argument & Computation 4 (2013) 46–63. doi:10.1080/19462166.2012.682656.

[6] R. Meester, The limits of bayesian thinking in court, Topics in Cognitive Science 12 (2020) 1205–1212. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12478. doi:https://doi.org/10.1111/tops.12478. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12478.

[7] R. J. Allen, M. S. Pardo, The problematic value of mathematical models of evidence., The Journal of Legal Studies 36 (2007) 107–40.

[8] A. Stein, Judicial fact-finding and the bayesian method: The case for deeper scepticism about their combination, The International Journal of Evidence Proof 1 (1996) 25–47. doi:10.1177/136571279600100103.

[9] R. Meester, L. Stevens, Bayesian reasoning and the prior in court: not legally normative but unavoidable., Law, Probability and Risk 23 (2024). doi:10.1093/lpr/mgae001.

[10] C. Dahlman, E. Kolflaath, The problem of the prior in criminal trials, in: Philosophical Foundations of Evidence Law, Oxford Academic, 2021.

[11] R. J. Allen, M. S. Pardo, Relative plausibility and its critics, The International Journal of Evidence Proof 23 (2019) 5–59.

[12] C. E. Berger, K. Slooten, The lr does not exist, Science % Justice 56 (2016) 388–391.

[13] D. J. N. Kristy A. Martire, Gary Edmond, B. R. Newell, On the likelihood of "encapsulating all uncertainty, Science % Justice 57 (2017) 76–79.

[14] A. Biedermann, T. Lau, Decisionalizing the problem of reliance on expert and machine evidence, Law, Probability and Risk 23 (2024).

[15] M. Sjerps, 2025. Personal communication.

[16] B. Y. Norman Fenton, Martin Neil, D. Lagnado, Analyzing the simonshaven case using bayesian networks., Topics in cognitive science 12 (2020) 1092–1114. doi:10.1111/tops.12417.

[17] L. V. Leeuwen, B. Verheij, A Comparison of Two Hybrid Methods for Analyzing Evidential Reasoning, in: Legal Knowledge and Information Systems, Madrid, 2019.

[18] B. Verheij, Proof with and without probabilities: Correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty, Artificial Intelligence and Law 25 (2017) 127–154. URL: http://link.springer.com/10.1007/s10506-017-9199-4. doi:10.1007/s10506-017-9199-4.
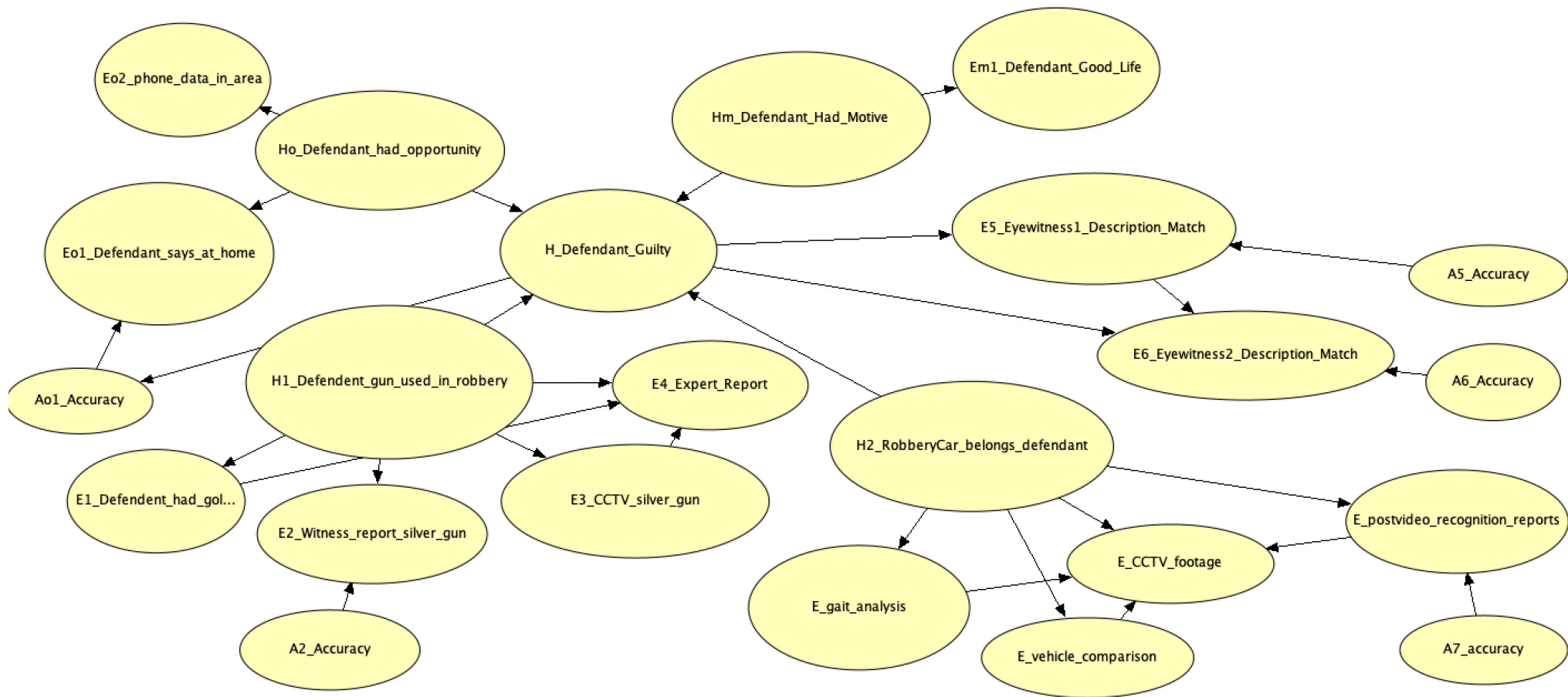
Appendix A: Modelling experience

Leya Lisa Hampson is a PhD student at the University of Groningen specializing in the application of Bayes in legal contexts. She has a background in mathematics and forensic science. She was
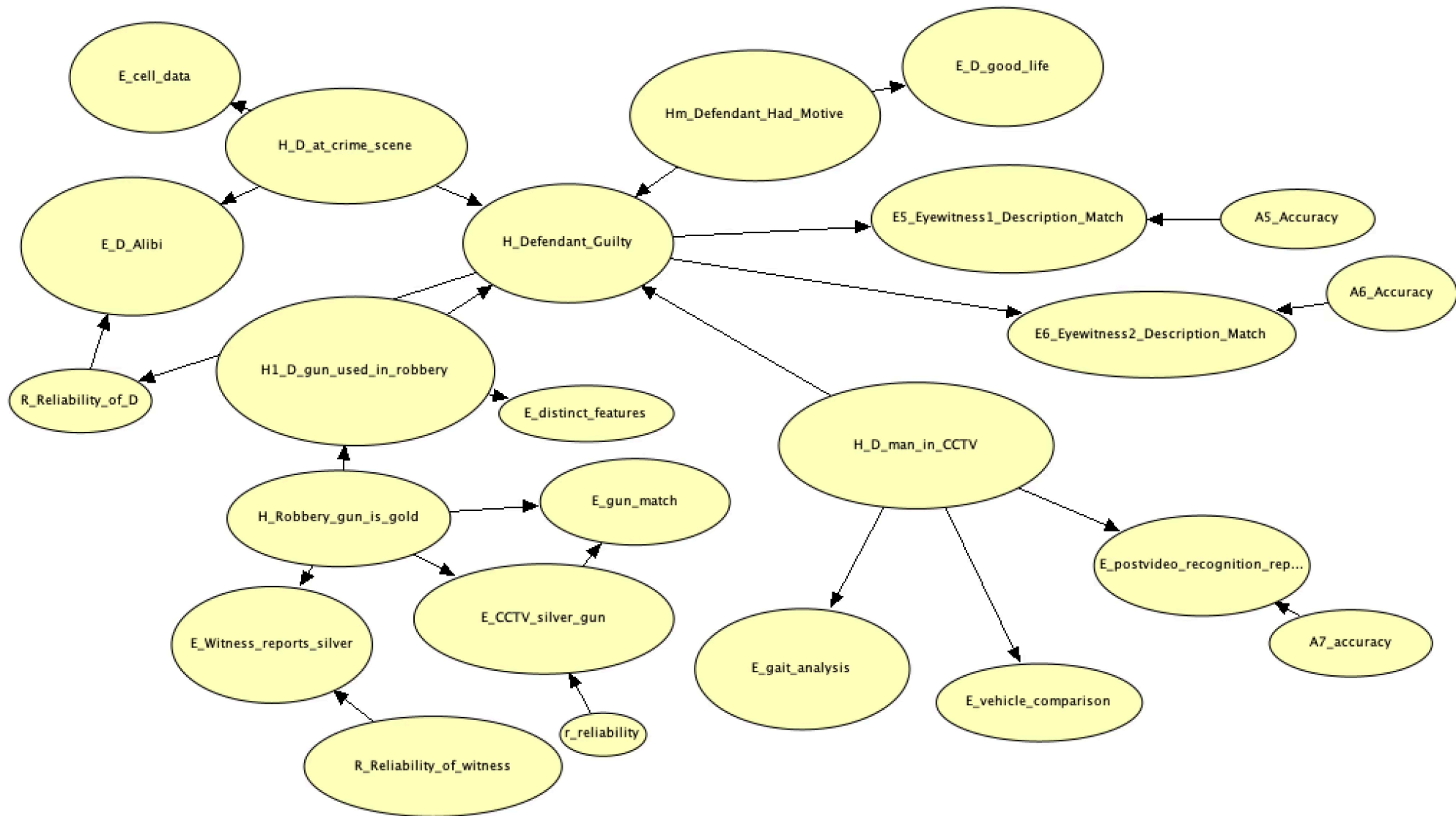
first introduced to Bayesian probability during her undergraduate studies in mathematics, where she developed a strong theoretical foundation in probabilistic reasoning. During her master's program, she completed the course 'Interpreting and Understanding Forensic Evidence', which focused on Bayesian modelling in legal contexts, based on Fenton et al. (2011) textbook. This knowledge was further applied and deepened in a course on digital evidence, in which Bayesian Networks were used to model evidential relationships in data-heavy cases. Over the course of her academic work, she has independently modelled approximately 6 complete legal cases using the AgenaRisk software. In addition, she has constructed numerous partial models and idiomatic structures, further reinforcing her practical experience with both qualitative and quantitative aspects of Bayesian modelling.

Ludi van Leeuwen is a PhD student at the University of Groningen specializing in evaluating Bayesian network models for legal evidential reasoning. She has a background in artificial intelligence and philosophy. During her Bachelor's degree, she compared different types of formalization of evidence by modelling a legal case [17] in both a Bayesian network, using the scenario idiom [4] as well as in a case model [18]. In her Master's in AI, her aim shifted to testing and evaluating idioms for Bayesian networks using artificial ground truths as groundings. In this project, continued into her PhD work, she has modelled 2 different (simplified) complete cases in Bayesian network and has implemented and evaluated numerous legal idiom structures. She usually models manually in Hugin, and generates Bayesian networks from artificial data using PyAgrum.

A. Full Bayesian Network Models

The full, enlarged Bayesian networks are provided below for reference.

width=!,height=!,pages=-, scale=0.9, pagecommand=, fitpaper=true
width=!,height=!,pages=-, scale=0.9, pagecommand=, fitpaper=true

Appendix C: Modeller 1: Variable and link modifications between the qualitative and quantitative networks

Table 7

Modeller 1: Variable Changes between Qualitative and Quantitative Models

Qualitative Network	Quantitative Network	Change	Related Evidence	Node Type
Ho_Defendant_had_opportunity	H_D_at_crime_scene	Variable name refined		Hypothesis
H2_RobberyCar_belongs_defendant	H_D_man_in_CCTV	Definition change		Hypothesis
	H_Robbery_gun_is_gold	Added	Gun	Hypothesis
Eo2_phone_data_in_area	E_cell_data	Variable name refined	Cell data	Evidence
E_Eo1_Defendant_says_at_home	E_D_Alibi	Variable name refined		Evidence
E_CCTV_footage		Removed		Evidence
E1_Defendant_had_gold_gun		Removed	Gun	Evidence
E_Expert_Report	E_gun_match	Variable name refined	Gun	Evidence
	E_distinct_features	Added	Gun	Evidence
Ao1_Accuracy	R_Reliability_of_D	Variable name refined		Reliability
A2_Accuracy	R_Reliability_of_witnesses	Variable name refined		Reliability
	R_reliability	Added		Reliability

Table 8
Modeller 1: Summary of edge changes between qualitative and quantitative BNs.

Qual. Parent	Qual. Child	Quant. Parent	Quant. Child	Change	Related to node change?
E5_Eyew1_Desc	E6_Eyew2_Desc	–	–	Removed	No
–	–	H1_gun_robbery	E_distinct_features	Added	Yes (Child node added)
H1_gun_robbery	E1_gold_gun	–	–	Removed	Yes (Child node removed)
E1_gold_gun	E4_expert	–	–	Removed	Yes (Parent node removed)
–	–	H_gun_gold	H1_gun_robbery	Added	Yes (Parent node added)
H1_gun_robbery	E2_silver_gun	H_gun_gold	H1_gun_robbery / E_silver_gun	Mediated	Yes (Sub-hypothesis added)
H1_gun_robbery	E4_expert	H_gun_gold	H1_gun_robbery / E_gun_match	Mediated	Yes (Sub-hypothesis added)
–	–	r_reliability	E_CCTV_silver_gun	Added	Yes (Parent node added)
H1_gun_robbery	E3_silver_gun	H_gun_gold	H1_gun_robbery / E3_silver_gun	Mediated	Yes (Sub-hypothesis added)
H2_RobberyCar	E_CCTV	–	–	Removed	Yes (Child node removed)
E_gait	E_CCTV	–	–	Removed	Yes (Child node removed)
E_postvideo_rec	E_CCTV	–	–	Removed	Yes (Child node removed)
E_vehicle_comparison	E_CCTV	–	–	Removed	Yes (Child node removed)

Appendix D: Modeller 2: Variable and link modifications between the qualitative and quantitative networks

Table 9
Modeller 2: Variable Changes between Qualitative and Quantitative Models

Qualitative Network	Quantitative Network	Change	Related Evidence	Node Type
Witness2recognizesCarAsDefen	Witness2RecognizesCar	Variable name refined	Car	Evidence
Witness2RecDefendant	Witness2RecognizesCar	Variable name refined	Car	Evidence
	DefendantReturnsToCar	Added	Car/scenario	Scenario
DefendantGun	DefendantHasCharAGun	Variable name refined	Gun	Scenario
GunFoundOnNightStand	GoldAGunFoundOnNightStand	Variable name refined	Gun	Evidence
LightingConditions	LightingConditionsAffordConfusion	Variable name refined	Gun	Reliability
Testimony	TestimonyPolice	Variable name refined	Gun	Evidence
	GunChars	Added	Gun	Scenario
RecontructtionGunLookedBothGoldSilver	ReconstructedGunDefendantShinesOppositeColor	Definition change	Gun	Scenario
GunisDefsGun	GunDefendantMatchesPerpetrator	Variable name refined	Gun	Scenario
	DefendantsOwnsGun	Added	Gun	Scenario
OtherCharacteristicsGunMatchesSuspect	CrimeSceneGunCharA	Variable name refined	Gun	Evidence
GunColorMatchesSuspectsGun	DefendantHasGoldCharAGun	Variable name refined	Gun	Scenario
DefendantCallsNearCrime	DefendantCallsFromNearCS	Variable name refined	Phonecall	Evidence
DefendantCallsNearHome	DefendantCallsNearHomeandCS	Variable name refined	Phonecall	Scenario
DefisPerp2		Removed	Scenario/identification	Hypothesis
CameraImageNoTattoo	CameraImageShowsTattoo	Definition change	Tattoo	Evidence
TestimonyP2	TestimonyP2SaysTattoo	Variable name refined	Tattoo	Evidence
TestimonyP1	TestimonyP1SaysTattoo	Variable name refined	Tattoo	Evidence
DefendantNoTattoo		Removed	Tattoo	Evidence
LaterMismembered	LaterStatementAccurate	Definition change	Witness 2	Reliability

Table 10
Modeller 2: Edge changes between qualitative and quantitative Bayesian networks.

Qualitative network		Quantitative ne...	
Parent	Child	Parent	Chi...
CarIsDefendants	PerpParksandExitsCar	–	–
–	–	DefendantParksandExitsCar	Car...
DefIsPerp1	PerpParksandExitsCar	–	–
PerpParksandExitsCar	PerpRobsShopThreatensWithGunandBreaksRegister	–	–
–	–	DefendantRobsShopThreatensetc	Def...
PerpThreatensWithGunandBreaksRegister	PerpReturnsToCarL1T2	–	–
–	–	DefendantRobsShopThreatensetc	Defe...
–	–	DefendantReturnsToCar	Perp...
–	–	GunChars	Gun...
–	–	CrimeSceneGunCharA	Gun...
–	–	GunColorCrimeScene	Gun...
DefIsPerp2	GunisDefsGun	GunDefendantMatchesPerpetrator	DefI...
GunColorCrimeScene	GunColorMatchesSuspectsGun	–	–
DefendantGun	GunFoundOnNightStand	–	–
–	–	DefendantOwnsGun	Defe...
–	–	DefendantHasGoldCharAGun	Gold...
DefendantGun	OtherCharacteristicsGunMatchesSuspect	–	–
OtherCharacteristicsGunMatchesSuspect	GunIsDefsGun	–	–

Appendix E: Modeller 1 completed qualitative reflection protocol

INDEPENDENT REFLECTION OF QUALITATIVE BNs

[Goal: Assessment of the use of qualitative BNs as a double-check for the court's reasoning]

Instruction: Please complete this checklist independently immediately after finishing the first modelling phase i.e. once you have constructed a qualitative BN for the case. This checklist is designed to support structured reflection of your model. The prompts serve as cognitive guides and highlight key issues for discussion. You are not expected to answer each question word-for-word or in list form. Instead, use them to organise your reflections and note any relevant insights, modelling decisions, uncertainties or concerns that arose during the modelling process.

1: ALIGNMENT	
(1) Does the BN fully represent all evidence cited in the court's verdict[1]?	Not entirely, I would have liked to incooperate more accuracy considerations of the evidence in more detail (i.e. the lighting, the witness relationship) rather than just summarised in one accuracy node (not possible due to Hugin limitations). I do believe however that all of the key evidence cited has been included. For example with eyewitness testimonies, one should consider veracity, competence and objectivity. Thus all evidence has been incoorperated but not all further evidence considerations.

---

[1] Any fact or piece of information cited by the court as part of its reasoning, whether directly discussed in terms of probative value or clearly presented as a basis for the final decision, is considered part of the evidential set to be modelled. Even if evidence is not explicitly assigned weight in the verdict, its inclusion indicates that it played a role in the court's assessment and should therefore be included in the model.

Appendix E: Modeller 1 completed qualitative reflection protocol

(2) Are (in)dependencies clearly defined?	As best as possible. As a lot of evidence is interconnected in this case, it is very important to be precise with (in)dependencies, however also very challenging. This was something I spent a lot of my time on. This is the part I am probably least confident in in my modelling, particularly regarding all evidence (potentially?) connected to the parking lot CCTV footage. I think the Dutch translation may have also complicated things here (unsure) however I sometimes found it difficult to interpret exactly which 'video evidence' for example they were referring to.
(3) Is the network sufficiently nuanced for the analysis it is meant to support? Are all chains of inference visible?	I do believe that all chains of inference are visible in the network, however, I would ideally liked to have expand the network even more to include detailed considerations of the accuracy of evidence items. For this specific analysis, as there is no strong probative evidence and a reliance on multiple witness reports, I honestly do not believe the network is sufficiently nuanced.
(4) Is the network unnecessary complex?	No, I do not believe so. I think visually the network could be more organised and structured but this does not affect actual complexity.

Appendix E: Modeller 1 completed qualitative reflection protocol

(5) Have alternative hypotheses been considered, including both the negation of the main hypothesis and any competing hypotheses?	Yes they have been considered, but I do not see any relevant alternative hypotheses.
(6) Does the structure mirror the logic of the court?	I think so.
EXTRA NOTES: Prior to begin of the experiment, I believed Hugin's 50 node constraint to be adequate for our modelling purposes. However, after modelling, I believe that it is not sufficient, as mentioned above, I would have liked to include more detailed considerations of accuracy.	
2: ERROR DETECTION	
(1) Does the BN implicitly expose any probabilistic fallacies? If so, which one(s)?	No.
(2) Are any logical fallacies or other reasoning errors (beyond calculation mistakes) present? If so, which one(s)?	I believe some accuracy considerations are missing in the court's logic but no specific errors.
(3) Is the direction of inference plausible i.e. are the directions of the links between the nodes consistent with the real-world causal relationships they represent?	Yes.
(4) Were any plausible alternative hypotheses ignored in the court's verdict?	No.
(5) Does building the BN expose ignored alternative hypotheses?	No.
EXTRA NOTES: N/A	
3: VALUE	

Appendix E: Modeller 1 completed qualitative reflection protocol

(1) How did building the qualitative structure increase your understanding of the case?	It forced me to think a lot deeper about the (possible) dependencies between the different pieces of evidence, thus greatly enhancing my understanding of the case. I do not believe just thinking about the case would have forced me to have such 'deep' internal debates on dependencies.
(3) Did building the qualitative BN allow you to identify any probabilistic fallacies?	**No.**
(4) Did building the qualitative BN allow you to identify other, non-probabilistic fallacies?	**No.**
(5) To what extent did Bayesian thinking assist the identification of missing evidence?	Bayesian thinking did assist the identification of missing evidence. I wonder if perhaps further CCTV footage could have explained/increased the match of the defendants car and the CCTV car, also is there no CCTV footage from the supermarket? I feel that it helped me identify some missing links that I would have liked to make/add.
(6) To what extent did Bayesian thinking assist the handling of (in)dependencies?	Very much so. Bayesian thinking forces you to think very deeply and intensely about the (in)dependencies between evidence, and visualising these via links in the BN allows for a great analysis. I think when I was just reading the case I was ignoring/neglecting a lot of dependencies that I later realised once modelling.

Appendix E: Modeller 1 completed qualitative reflection protocol

(7) Did building the qualitative model help you expose any jumps in reasoning? If so, did this reveal any (un)acceptable implicit assumptions?	I find the car match to the defendants a bit of a jump in reasoning. Additional evidence could help expose whether this was a harmful jump or not. I further think that the identification of the gun poses a gap in reasoning, I believe more information on the frequency of this type of gun (rather than just CCTV match statements) is needed to make the assumption that this gun is his.
EXTRA NOTES:	
4: QUANTIFICATION	
(1) Which parts of the structure would most benefit from quantification?	None of the evidence in the case has very strong probative value individually. Therefore, I honestly think this is a case where the full net would highly benefit from quantification. However, I think it is most important/most interesting to me to quantify the many witness/eyewitness statements, to see how large their impact is on the final verdict. Also the subhypothesis that the car belonged to the defendant would be very interesting to see in numbers, with an in depth analysis of how many such cars there are in the area.
(2) Which assumptions do you expect the quantitative structure to confirm, challenge or clarify?	I hope for the quantification of the net to further clarify the match of the car, as this currently seems a bit weak to me. The defence' statement of a missing

Appendix E: Modeller 1 completed qualitative reflection protocol

	motive will be confirmed by the net I assume, as there is no strong evidence to say otherwise. I believe the quantification may challenge the reliability of the many witness statements more, as specific considerations on the accuracy of each are factored in in the qualitative net, of which their true value and impact will be exposed once numbers are added.
(3) Is overall quantification necessary or useful in evaluating whether this model meets the BARD threshold?	Yes, definitely.
EXTRA NOTES:	

Appendix F: Modeller 2 completed qualitative reflection protocol

INDEPENDENT REFLECTION OF QUALITATIVE BNs

[Goal: Assessment of the use of qualitative BNs as a double-check for the court's reasoning]

Instruction: Please complete this checklist independently immediately after finishing the first modelling phase i.e. once you have constructed a qualitative BN for the case. This checklist is designed to support structured reflection of your model. The prompts serve as cognitive guides and highlight key issues for discussion. You are not expected to answer each question word-for-word or in list form. Instead, use them to organise your reflections and note any relevant insights, modelling decisions, uncertainties or concerns that arose during the modelling process.

1: ALIGNMENT	
(1) Does the BN fully represent all evidence cited in the court's verdict?	Most of it. Evidence that was not modelled: - The details of the car (matching wheels/trekhaak/interior colors) - Footage of the defendant at his house, establishing the car exit - Alternative explanations for the suspect's strange gait/exit are not modelled. - Footage of defendant in car at gas station - Details of personal identification (length, hair, clothing, age) - Precise specification of gun details except colors - Precise specification of gait analysis (also not in verdict) - How the gun was identified in the camera images
(2) Are (in)dependencies clearly defined?	**Yes, but they are difficult to think about without entering the CPTs**

Appendix F: Modeller 2 completed qualitative reflection protocol

(3) Are all chains of inference visible?	**Reliability of witnesses not always modelled explicitly, chain of inference from perp to defendant is visible. Some nodes have been collated into a single node (PerpParksExitsCar, PerpRobsShop).**
(4) Have alternative hypotheses been considered, including both the negation of the main hypothesis and any competing hypotheses?	**Yes, orange nodes for alternative explanations (both on the prosecution side and on the defense side), but not represented in a single node/single alternative story**
(5) Does the structure mirror the logic of the court?	**The evidence is considered in separate clusters, from left to right there is a "timeline". The identification of the suspect as the defendant is made explicit. I think to a large extent the structure mirrors that of the court.**
EXTRA NOTES:	
2 : VALUE	
(1) Did building the qualitative structure increase your understanding of the case?	**Yes, but I was confused sometimes about how to model things (such as "identification"). There seemed to never be a doubt about the facts that occurred, only who did them. The BN identifies the "source" of uncertainty.**

Appendix F: Modeller 2 completed qualitative reflection protocol

(2) Did you detect issues you may have missed using text-based reasoning alone?	**Yes, it is not made explicit if the phone signal of the defendant near the crime scene could also be due to that being near his house.** **It is not clear how often these similar cars occur, like how many cars were found in the list to be investigated?**
(3) Does the building the qualitative BN allow for the identification of probabilistic fallacies?	**I think something went wrong in the case with the gait/exit likelihoods, but I'm not sure if I modelled that correctly in the BN either. I will need to add numbers for that.** **Also, independent witnesses and double-counting seems important here.**
(4) Does building the qualitative BN allow for the identification of other, non-probabilistic fallacies?	**Not sure**
(5) To what extent did Bayesian thinking assist the identification of missing evidence?	**There seemed to be missing evidence for the alternative explanations, also the reliability of witness 2, and the whole process seems to hinge on the suspect driving this car & then finding the similar gun.**
(6) To what extent did Bayesian thinking assist the handling of (in)dependencies?	**I think something is wrong with the gait/movement testimony re:independence.**

(7) Did building the qualitative model expose any jumps in reasoning? If so, did this reveal any (un)acceptable implicit assumptions?	**Doesn't seem to have established that the defendant actually physically looked like the suspect, apart from that they have similar heights and accents. Also, did not consider the alternative explanation of his phone sending to the mast if that was also where he lived.** **Also: jumps to assume that the suspect running to and from the car was the same as the man in the store, even though that's not justified in the verdict.**
EXTRA NOTES:	
3: ERROR DETECTION	
(1) Does the BN implicitly expose any probabilistic fallacies? If so, which one(s)?	**Can't see that now. Maybe the dependence of the gait/exit (as I can imagine), or a reference class problem for the car…, or neglecting the alternative hypothesis that the phone is near his home.**
(2) Are any logical fallacies or other reasoning errors (beyond calculation mistakes) present? If so, which one(s)?	**Not sure.**
(3) Is the direction of inference plausible i.e. are the directions of the links between the nodes consistent with the real-world causal relationships they represent?	**Yes, I mostly used evidence-idiom constructions with some abstract node relating to 'identification'**
(4) Are any plausible alternative hypotheses ignored?	**Can't say without quantification.**

Appendix F: Modeller 2 completed qualitative reflection protocol

(5) Does building the BN expose ignored alternative hypotheses?	**Call near home,**
EXTRA NOTES:	
4: QUANTIFICATION	
(1) Which parts of the structure would most benefit from quantification?	**All of it!** **I find it hard to judge whether the structure is correct without seeing if setting some evidence aligns with what I think it should do.**
(2) Which assumptions do you expect the quantitative structure to confirm, challenge or clarify?	**I hope it will give me insight into to what extent the evidence is strong enough to carry the "identification" of the perpetrator as the defendant.**
(3) Is overall quantification necessary or useful in evaluating whether this model meets the BARD threshold?	**YES**
EXTRA NOTES:	