A normative model of explanation for binary classification legal AI and its implementation on causal explanations of Answer Set Programming

Evan Iatrou[1]

[1]*Institute of Logic, Language and Computation (University of Amsterdam), Science Park 107, 1098 XG Amsterdam, The Netherlands*

Abstract

In this paper, I provide a *normative* model of explanation for the output of AI algorithms used in legal practice. I focus on *binary classification* algorithms due to their extensive use in the field. In the last part of the paper, I examine the model's compatibility with *causal* explanations provided by *Answer Set Programming* (*ASP*) causal models.

The motivation for proposing this model is the *necessity* for providing explanations for the output of legal AI. From the multiplicity of arguments supporting that necessity, the proposed model addresses the argument that legal AI's output should be *objectionable*. That can be achieved only if the explanation of the output has a *form* that makes it *amenable* to evaluation by legal practitioners. Hence, I firstly provide a normative model for the explanations used by legal practitioners in their practice (CLM_{LP}) and then I provide the normative model for the explanations of legal AI's outputs (EXP_{BC}) that I base on CLP_{LP}. CLP_{LP} in its turn is based on the *Classical Model of Science* (*CMS*) which is the normative model of explanations that every "*proper*" science should follow according to philosophers throughout history.

Following the introduction of EXP_{BC}, I propose three *degrees of explainability* regarding binary classification explanations according to their fidelity to EXP_{BC}. I further argue that *machine learning* can not satisfy even the lowest degree of explainability, while *rule-based* AI - like ASP-based AI - can satisfy the highest degree. Concluding, I propose an ASP methodology in-progress that can use EXP_{BC} to provide *causal* explanations. In the proposed ASP methodology, I am using *causal graphs* as models of causal inference as well as a metaphysically neutral *interventionist* account of causation.

CMS_{LP} and EXP_{BC} are normative models that are based on the derivation relation of *subsumptive-deductive* inference among *norms* and *propositions*. On the other hand, the proposed ASP methodology is based on *causal* - and hence non-subsumptive-deductive - relations among norms and propositions that *supervene* on the subsumptive-deductive ones. Consequently, the motivation behind proposing this specific ASP methodology is to establish a precedent for a *unification* of different types of explanations (e.g., deductive and causal explanations) as well as to bridge the gaps among *computational modelling*, *actual practice*, and the *philosophical underpinnings* of a domain of expertise (*law* in this case).

Keywords

Classical Model of Science (CMS), legal XAI, degrees of explainability, causal graphs, Answer Set Programming (ASP),

ICLP Workshops 2022: 15th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP), July 31, 2022, Haifa, Israel.

evangelos.iat@gmail.com (E. Iatrou)
https://www.illc.uva.nl/People/person/5123/Evangelos-Iatrou (E. Iatrou)
0000-0002-9866-1477 (E. Iatrou)

1. Introduction

[1]AI has already multiple applications in legal practice. Public judicial institutions employ AI to assess the possibility of the defendant recidivating [1]. Private companies use AI to review thousands of documents to determine which ones are relevant to a particular case [2]. There are even "*robot lawyers*" like the *DoNotPay* app;[2] an initiative that started by contesting parking tickets and now it has expanded to other diverse services - from landlord protection to canceling a *Disney+* subscription.

A large amount of the legal AI literature is devoted on predicting the outcome of a case (e.g., whether the Court will rule in favour of the prosecution [3, 4, 5]) and whether a legal document belongs to a certain category (e.g., category$_1$="*The document is relevant to a pending case.*", category$_2$="*The document is not relevant to a pending case.*" [6, 7, 8]). There are also many cases of AI that are not usually considered by the general public as legal AI while they should. Such an example with significant impact on the daily lives of billions of people are *upload filters* - i.e., AI that "*substitutes*" legal institutions (e.g., courts) by classifying data as (un)lawful [9]. All foregoing examples are essentially *binary classification* tasks: the defendant *is* or is *not* guilty, a document is or is *not* relevant to a pending case, an uploaded post *contains* or it *does not* contain unlawful hate speech. Considering its wide range of applications, the rest of the paper will be devoted to binary classification legal AI.

Legal experts advocate that AI used in legal practice *must* provide an *explanation* for its output. One of the central arguments is that the outcome of such an algorithm should be *objectionable* [10]. That can be achieved only if the explanation of the output has a *form* that makes it *amenable* to evaluation by legal practitioners. As a result, any *normative* model of explanations of legal AI's output should be constructed based on a normative model of explanations used in legal practice so as to *justify* its normativity. That also means that any normative model of explanations of legal AI's output should be constructed *prior* to its implementation to actual AI models since the latter should follow the former. This is why, in this paper, I do not propose a normative model derived from the current state-of-the-art legal AI, but on the contrary, it is a model that the state-of-the-art legal AI *should* follow. The characterisation "*state-of-the-art*" itself can not be attributed without a *prior* listing of all the normative explainability requirements from the prospective of the legal practitioner that will actually use the proposed AI model.

The paper is structured in 5 sections with §1 being the introduction and §5 the conclusion. In §2, I construct a normative model of legal practice. Specifically, in §2.1, I clarify the meaning of the term "*legal practice*" in the context of this paper. In §2.2, I present the Classical Model of Science (CMS), i.e., a normative model of every "proper science" according to philosophers throughout history. Finally, in §2.3, I apply the CMS to legal practice in order to construct a normative model of legal practice that I name CMS_{LP}.

In §3.1, I use the normative model of legal practice from §2.3 to finally construct a normative model of explanation for the output of binary classification legal AI. I name this model EXP_{BC}. In §3.2, I use this model to classify the explanations provided by current legal XAI methods in three categories. Those categories are ordered in *degrees of explainability*. I argue that *rule-based*

[1]**Original submission**

[2]https://donotpay.com/ (last visit: 10/04/2022)

AI can satisfy the highest degree of explainability while *machine learning* can not even satisfy the lowest degree.

So far, all proposed normative models (CMS_{LP}, EXP_{BC}) are based on *subsumptive-deductive* inference - the principal reasoning method used in legal practice. In §4 - the final section of the paper, I argue how *Answer Set Programming* (*ASP*) - as a paradigmatic case of rule-based programming - can be employed to: (i) model *causal inference* based on subsumptive-deductive inference;[3] (ii) use causal inference's *supervenience* on subsumption-deduction to satisfy the highest degree of explainability presented in §3.2. The motivation for this approach to causal explanations is to establish a precedent for a *unification* of types of explanations (e.g., deductive and causal explanations) and to bridge the gaps among *computational modelling*, *actual practice*, and the *philosophical underpinnings* of a discipline (the discipline of *law* in this case). Due to the latter motivation, the paper is written in a way that makes it accessible to all addressed audiences. The endeavour of addressing multiple disciplines is an emerging challenge for the practical use of AI and *ergo*, a secondary issue this paper tries to indirectly tackle not explicitly but *by example*. E.g., by avoiding unnecessary AI or philosophical jargon and by providing examples and intuitions where necessary all the while retaining the content's quality.

2. A normative model of legal practice: the Classical Model of Science

2.1. Legal practice

Before constructing a normative model of *legal practice*, I have to clarify what legal practice *is* in the context of this paper. By "*legal practice*", I mean the totality of the *activities* of the legal experts that are *authorised* to take part in the process of deciding a judgement (e.g., the lawyers of the defense and the prosecution, the judges of the authorised court). Those activities can be for instance the arguments the defence makes during a trial and all of the defence's preparatory work before the trial. The authorisation of the legal experts is stipulated by law. E.g., the *European Convention of Human Rights (the Convention)* stipulates that the authorised court for deciding whether there has been a violation of the Convention's articles is the *European Court of Human Rights (ECtHR)*.[4]

I borrow this construal of legal practice from [11]. My motivation is that according to this construal, the produced results of legal practice are *legally binding*. And *being legally binding* will be a *necessary property* for the conception of *truth* I will employ in the proposed normative model of legal practice.

2.2. The Classical Model of Science (CMS)

The normative model of legal practice I will provide will be based on the *Classical Model of Science (CMS)*. CMS is a posterior reconstruction of *how* the *ideal scientific explanation* for every "*proper science*" should look like according to philosophers throughout history [12]. One may

[3] My choice of causal inference models is *causal graphs* based on a metaphysically neutral *interventionist* account of causation.

[4] *Article 19* of the Convention.

object to my decision to provide a normative model of legal practice using the CMS since the latter is about *sciences* and not about a *practice*. However, even if legal practice is not a science - whatever that means - my arguments supporting that legal practice fits into the CMS will still hold since *none* of those arguments is based on the premiss that legal practice is a science.

Let's move to the presentation of the CMS. I will provide a less detailed account of the CMS which I will call CMS_{sum}. For the construction of EXP_{BC}, there is no need to argue how legal practice fits into the more detailed version of CMS_{sum} found in [12]. Specifically, EXP_{BC} will be a *minimal* model. By "*minimal*", I mean that one can further expand and/or precisify parameters of that model[5] using the more detailed version of the CMS_{sum}.[6] However, for that to be done properly, I would have had to exceed the word limit.

(SUM.1) CMS_{sum} is a system S of propositions, concepts (or terms) which satisfy clauses (SUM.2) to (SUM.6).

(SUM.2) CMS_{sum} has a *domain* D. The propositions, concepts and terms of S are about the entities of D.

(SUM.3) The propositions form a *hierarchy* H_P based on their *derivation*: the propositions situated higher in the hierarchy are derived from the propositions situated lower.

(SUM.4) The concepts and the terms form hierarchies H_C and H_T respectively based on their *derivation*: the constituents of a hierarchy situated higher in the hierarchy are derived from the constituents situated lower.

(SUM.5) All propositions of S are *true*.

(SUM.6) The domain experts *know* that the propositions of S are true and they have *adequate knowledge* of all concepts and terms of S. For each hierarchy $H_{i,i\in\{P,C,T\}}$, knowledge about one of its constituents is *justified* from knowledge about other constituents situated lower in that hierarchy.

2.3. Applying CMS_{sum} to legal practice

As noted by [12], for each application of the CMS, CMS's parameters may have to be modified to fit the particularities of that application. Below, I perform such modifications to apply CMS_{sum} to legal practice. In §2.4, I sum up the outcome of those modifications to introduce a *normative model of legal practice* which I will name CMS_{LP}.

• Applying (SUM.3): Legal reasoning can be construed as rule-based reasoning [13]. One way to perform such a construal is to view a legal inference as an inference whose premisses consist of *particular facts* and *general rules* [14] like the following example borrowed from [15]:

[5] By "*parameters*", I refer to all the *objects* and *relations* that the CMS refers to: propositions, concepts, terms, derivation relations, the domain D, the concepts of *truth* and *knowledge*, etc.

[6] An example of such a pacification is to narrow down the derivation relations among propositions to those of *deduction* and *causality*. That way we exclude all other derivation relation like reasoning by *analogy* or metaphysical *grounding*.

Example 1.

(n_1) If x lives in Italy for more than 183 consecutive days over a 12-month period, then x is obliged to pay taxes in Italy on their worldwide income. (general rule)

(p_1) Alice lives in Italy for more than 183 consecutive days over a 12-month period. (particular fact)

(n_2) Alice is obliged to pay taxes in Italy on their worldwide income. (conclusion)

A particular fact is a *proposition* while a (general) rule is a *norm*. Intuitively, the difference is that propositions *describe* which state of affairs *is* the case (e.g., "*Alice* **is** paying taxes.") while norms *stipulate* which state of affairs *should be* the case (e.g., "*Alice* ***should pay*** *taxes.*"). Since legal inferences - like Example 1 - contain both propositions and norms, we can not explain the inferences' conclusions without appealing to the referred norms. Consequently, CMS_{LP} has to contain norms apart from propositions. Moreover, Example 1's conclusion is also a rule - not a general, but a *particular* one - and hence a norm. In other words, we can *derive* new norms from other norms and propositions. Consequently, the hierarchy H_P in CMS_{LP} contains both norms and propositions which are ordered based on *derivation* relations; H_P's constituents situated higher are *derived* from those situated lower. E.g., norm n_2 is situated higher in H_P than norm n_1 and proposition p_1. The derivation relations whose conclusions are norms are those of *subsumptive-deductive* inferences: the particular fact (2) is *subsumed* by rule (1) and via *deduction* the conclusion (3) follows.

In the context of this paper, the *output* of a binary classification algorithm is essentially a *judgement*. E.g., a court's judgement about whether there has been a violation of article X or not. As such, a judgement is the *conclusion* a legal inference. In Example 1, that conclusion has the form of a *norm*. However, we can always construe a judgement as a *proposition* p and not as a *norm* n. Specifically, $p := p_n$ is a *descriptive meta-analysis* of n whose existence is "*parasitical*" to that of n [16]. E.g, instead of $n =$"*Alice should pay taxes.*", we can always say $p_n =$"*Alice is obliged by law to pay taxes.*". Having said that, whenever I discuss about a judgement as part of legal inference (e.g., Example 1) I will construe it as a *norm* unless explicitly saying otherwise.

But what are the norms of H_P? In §2.1, I construed legal practice as "*the totality of the activities of the legal experts that are authorised to take part in the process of deciding a judgement*". The foregoing *authorisation* of the legal experts is authorised by the laws of a *specific* legal system. E.g., a court's judgement about Alice's taxes to the Italian State will be based on the norms of the Italian legal system. At the same time, if Alice's activities lie outside of the jurisdiction of French law, then the norms of the French legal system are not applicable to them. Consequently, different legal systems may regulate different sets of entities D (*domains*). As a result, the practice in each domain will be different. In other words, to every legal system corresponds a different CMS_{LP} whose norms are: (i) those that make up the legal system; (ii) those that can be derived by the norms of (i) in a similar manner with Example 1.

• Applying (SUM.4): *Subsumption* is essentially a *decision* performed by the authorised legal experts; the experts decide *inter alia* whether *terms* appearing in the particular fact fall under *concepts* appearing in a rule. E.g., whether the term "*Alice*" belongs to (*is subsumed by*) the concept of "*living in Italy for more than 183 consecutive days over a 12-month period*"[14]. Consequently, CSM_{LP} includes hierarchies H_C and H_T of the *concepts* and *terms* participating in the subsumption.

For the application of (SUM.5) to legal practice described right afterwards, it is important to highlight that the concepts of H_C are *interpretive concepts*: an interpretive concept of a domain D is a concept for which the domain experts can *not* concede on a specific list of criteria about whether an entity of D belongs to that concept. The concepts for which domain experts *do* concede on such criteria are called *criterial concepts*. For instance, the domain experts of biology agree that there is a a unique criterion for whether an entity belongs to the concept of *tiger*: the entity has *tiger DNA*. In law, there is no such thing as a DNA of *justice*, *freedom*, *dignity* and the rest of legal concepts. To make things worse, the decision of whether an entity belongs to a legal concept is influenced by the background beliefs (ethical and political) of the authorised legal practitioners [17]. Hence, the extensions of the concepts of H_C are decided on *subjective criteria*. Taking that into consideration, one could argue that CMS_{sum} is *not* applicable to legal practice. Specifically, according to [12], the hierarchy H_C reflects "*real or objective grounds (aitiai) of things*" (*ordo essendi*). Therefore, since interpretive concepts' extension are depended on subjective criteria they can not be grounded on "*real or objective*" *aitiai*.

To respond to that objection, I adopt the construal of legal interpretive concepts found in [18]. In brief, their argument is that in each *tradition* of legal practice (e.g., in the tradition of Italian legal practice) at a given point in time a *unique* conception C of a legal concept emerges. The legal practitioners *interpret* that concept based on their background beliefs. Then, via their *disagreement* which is performed according to the methods and customs of that tradition, they *concede* at a decision of whether an entity belongs to C. Ideally, their *consensus* coincides with the *ordo essendi*. Note that the decision about the extension of a legal concept is vital for legal practice since the *truthfulness* of a *subsumption* depends on that decision.

• Applying (SUM.5): The *propositions* of CMS_{LP} refer to *state of affairs* among CMS_{LP}'s concepts and terms. Since there is an *ordo essendi* about these concepts and terms, there is also an *ordo essendi* about their state of affairs and hence, *all propositions* of CMS_{LP} are *true*.

Regarding the *truthfulness* of CMS_{LP}'s *norms*, we have to deal with the objection that norms can *not* take truth values since they do not attempt to describe actual states of affairs. From all available responses to that problem, I adopt the response that a norm is true in a given normative system *in virtue of* that system including that norm.[7] E.g., an Italian law is true in the Italian jurisdiction in virtue of belonging to the Italian legal system. Apart from the norms that are included in a normative system, we can infer new norms *via* subsumptive-deductive inference as is done in Example 1. Those norms are true *in virtue of* a truthful subsumption-deduction.

I will notate the conception of truth I have used so far as truth_1. It is the truth reflected in the *ordo essendi*. However, next to truth_1, we need to include another notion of truth: truth_2. Specifically, as elaborated in §2.1, a judgement n in the context of legal practice is decided by a group of *authorised* domain experts. Hence, it may be the case that the experts' judgement does *not* correspond to the *ordo essendi*. A usual cause of such differentiations is the mismatch between the consensus of the authorised legal experts on the *extension* of a legal concept and the *ordo essendi*. Despite that, judgement n is *still* established as true *in virtue of* the authority of the authorised legal experts since their judgements are *legally binding* [16]. This is truth_2.

[7] To argue in favour of that response, I would have to exceed both the scope and the word limit of this paper. Hence, I take it as a given.

In contrast with legal practice, in the practice of empirical sciences, no group of experts has the authority to establish truths; the domain experts do not hold any authority over the physical world [16]. For instance, Alice has cancer independently of the experts' conclusion, while O. J. is guilty in virtue of the experts' conclusion (truth_2) and independently of whether he indeed killed his wife (truth_1). There are of course institutional ways to challenge truth_2 which are still though appeals to another group of authorised experts - like appealing to a higher court. On the contrary, in empirical sciences, a conclusion can *always* be challenged, and *by any one* of the domain experts [16].

In case that the legal concepts of CMS_{LP} have been interpreted by the authorised legal experts (truth_2) in a different way than the *ordo essendi* (truth_1), they may also compose a *different* hierarchy than the hierarchy H_C of the *ordo essendi*. I will notate that hierarchy as $H_{C,2}$. Similarly, since the concepts of $H_{C,2}$ play a decisive role in the subsumption of particular facts (propositions) by general rules (norms) in legal inferences, those propositions and norms may compose a different hierarchy than H_P. I will notate that hierarchy as $H_{P,2}$. For instance, it may be the case that according to the *ordo essendi*, a particular fact f is subsumed by a norm n_1 allowing us to infer another norm n_2. In that case, n_2 is situated higher in H_P than n_1 and f. However, if the experts interpret the concepts of n_1 differently than the *ordo essendi*, it may be the case that f can not be subsumed by n_1 and subsequently, we can no longer infer n_2. In that case, n_2 is *not* situated higher than n_1 and f.

• Applying (SUM.6): Since legal practitioners *must* accept the judgements (conclusions) of the authorised domain experts any conception of *knowledge* and *justification* must be based on truth_2.

• Applying (SUM.2): As mentioned in ¶4 of (SUM.3)'s application to legal practice, CMS_{LP}'s domain D consists of all the entities that CMS_{LP}'s norms regulate. The concepts, terms and propositions of CMS_{LP} are about the entities of D.

2.4. CMS_{LP}: a normative model of legal practice

In this subsection, I introduce *CMS_{LP}*, i.e., a *normative model of legal practice*. It is the result of the application of CMS_{sum} to legal practice as elaborated in §2.3.

(LP.1) CMS_{LP} is a system S of propositions (*particular facts*), norms (*general rules*), concepts and terms which satisfy clauses (LP.2) to (LP.6).

(LP.2) CMS_{LP}'s domain D consists of all the entities that the norms of CMS_{LP} regulate. The propositions, concepts and terms of CMS_{LP} are about the entities of D.

(LP.3) The propositions and norms of CMS_{LP} form a hierarchy H_P based on their *derivation*: the propositions and norms situated higher in the hierarchy are derived from the propositions and norm situated lower. The derivation of norms from other norms and propositions has the form of *subsumptive-deductive* inference.

(LP.4) The concepts and the terms form hierarchies H_C and H_T respectively based on their *derivation*: the constituents of a hierarchy situated higher in the hierarchy are derived from the constituents situated lower. The concepts are *interpretive* and *objective* (i.e., there is an *ordo essendi*).

(LP.5) There are two notions of truth: (i) truth_1 that reflects the *ordo essendi*; (ii) truth_2 which is established by a group of authorised domain experts. The authority of the authorised experts is authorised by the norms of H_P. In legal practice, the truthfulness of a judgement is true in the sense of true_2.[8] truth_2 induces new hierarchies $H_{P,2}$, $H_{C,2}$. truth_1 coincides with truth_2 *only if* $H_{C,2}$ coincides with H_C.

(LP.6) The domain experts *know* that the propositions and norms of $H_{P,2}$ are true (truth_2) and they have *adequate knowledge* of all concepts and terms of $H_{C,2}$ and H_T respectively. For each hierarchy $H_{P,2}$, $H_{C,2}$ and H_T, knowledge about one of its constituents is *justified* from knowledge about other constituents situated lower in that hierarchy.

3. A normative model of the explainability of upload filters

3.1. The model

Assume a set of norms $\mathcal{N}$ that are legally binding for a set of data D (e.g., hateful tweets). Then, based on the norms $\mathcal{N}$, D is segmented into two *disjoint* sets: $D = D_{\mathcal{N}} \cup D_{\overline{\mathcal{N}}}$. Data d are elements of $D_{\mathcal{N}}$ *if and only if* they do *not* violate the norms of $\mathcal{N}$. Equivalently, d are elements of $D_{\overline{\mathcal{N}}}$ *if and only if* they *do* violate the norms of $\mathcal{N}$. Since $\mathcal{N}$ are norms that are *legally* binding for D, any judgement on the violation of $\mathcal{N}$ by D's data is true *if and only if* it coincides with the judgments of the *authorised* legal practitioners. Since I want to provide a *normative* model and not a *descriptive* one, I assume that there is an algorithm $\mathcal{F}$ that segments (or *filters*) D to the two disjoint sets $D = D_{\mathcal{N}} \cup D_{\overline{\mathcal{N}}}$ with 100% accuracy (*ideal - normative* case). Propositions of the form $p := d \in D_i$ - where $i \in \{\mathcal{N}, \overline{\mathcal{N}}\}$ - are *judgements* for the legality of d. Since p represents a judgement, we can substitute it by a norm n_p that represents the same judgement.[9]

Since n_p is a judgement it is part of legal practice. Hence, its truthfulness corresponds to truth_2. In this *normative* context, truth_2 is the judgment to which the authorised legal experts *would have conceded if the judgement had been decided by them* and not by $\mathcal{F}$. Moreover, since n_p is part of the legal practice, its explanation *should* be based on the *normative* model CMS_{LP}. According to CMS_{LP}, the truthfulness of n_p is *grounded* on the hierarchy $H_{P,2}$: there are propositions and norms situated lower than n_p in the hierarchy $H_{P,2}$ such that n_p is derived from them *via* subsumption-deduction. Consequently, an *explanation* of n_p *should* consist of the *subsumptive-deductive arguments* whose conclusion is n_p. Finally, the user whose data are filtered *knows* the normative explanation of $\mathcal{F}$'s output when they *know* those subsumptive-deductive arguments.

(BC.1) Assume a dataset D, a set of norms $\mathcal{N}$ which are legally binding for D, and an algorithm $\mathcal{F}$ that segments (*filters*) D to $D = D_{\mathcal{N}} \cup D_{\overline{\mathcal{N}}}$ with 100% accuracy. Then, EXP_{BC} is the CMS_{LP} whose norms include $\mathcal{N}$ and the norms n_p, where $p := d \in D_i$ and $i \in \{\mathcal{N}, \overline{\mathcal{N}}\}$.

(BC.2) EXP_{BC}'s domain is the dataset D.

[8] Since CMS_{LP} is *normative*, one could argue that $true_2$ should coincide with true_1; that is the *ideal (normative)* case. However, since I want to provide a model of legal *practice*, I want it to be *pragmatically* useful. I.e., I want a model that is applicable to every group of *actual* legal practitioners. Therefore, the ideal state of affairs in every trial is the ideal state of affairs *relevant* to the capabilities of the authorised legal practitioners. Consequently, it may be the case that even in their ideal (normative) performance, those practitioners are not capable of discovering truth_1.

[9] *See* ¶3 of (SUM.3)'s application to legal practice in §2.3.

(BC.3) The truthfulness of the partition $D = D_{\mathcal{N}} \cup D_{\overline{\mathcal{N}}}$ is that of truth_2.

(BC.4) An explanation of a judgement n_p consists of:

(BC.4.a) all the propositions and norms that are the premisses of the subsumptive-deductive inference whose conclusion is n_p

(BC.4.b) the way those propositions and norms are structured (*argumentative structure*) to form the subsumptive-deductive inference of (BC.4.a).

(BC.5) A user whose data are filtered by $\mathcal{F}$ *knows*:

(BC.5.a) the propositions and norms described in (BC.4.a)

(BC.5.b) the argumentative structure described in (BC.4.b).

3.2. EXP_{BC} and the current state of AI: machine learning *vs.* rule-based programming

EXP_{BC} is compatible with the the review of the current forms of explanation for the output of legal AI found in [10]. Based on that review, we can classify different explainable legal AI (or legal *XAI*) models into three categories of different *degree* of explainability:[10]

(*degree 1*) Explanation that consist only of a subset of the *propositions* of (BC.4.a).

(*degree 2*) Explanation that consist only of a subset of the *propositions* and *norms* of (BC.4.a).

(*degree 3*) Explanation that consists of *parts of* the argumentative structure described in (BC.4.b).

From the above classification, we can infer that the content of explanations of *degree 1* is contained in the content of explanations of *degree 2* whose content is contained in the content of explanations of *degree 3*. Hence, explanations of *degree 3* contain the same and more *information* of explanations of *degree 2* which contain the same and more *information* than explanations of *degree 3*. The *ideal* would be for the XAI model to provide maximum information in its explanations. I.e., provide an explanation of *degree 3* that contains *all* the propositions and norms of (BC.4.a) and their *complete* argumentative structure described in (BC.4.b).

3.2.1. EXP_{BC} and machine learning

Although I find Adrien et al.'s 2021 clustering of types of explanation quite useful, there is an important divergence between my interpretation of that classification and theirs. In contrast with them, I do not consider *machine learning (ML)* algorithms capable of satisfying *any* of the three degrees of explainability.

Take for instance the example of [3] which [10] present as a state-of-the-art case of explanation of *degree 3*. Their model is a binary classification ML algorithm. The designing of binary classification ML algorithms consists of two phases: (a) *training* phase; (b) *testing* phase. During the training phase, the model is "*fed*" with data from two categories and it extracts patterns that appear with *high frequency* in each category. During the testing phase, the model attempts to identify such patterns in new data and it classifies them accordingly. In case that the classification of the testing phase has low accuracy, the model is re-trained and so on [19].

[10] The following is my *reconstruction* of Adrien et al.'s 2021 classification.

The binary classification algorithm that [3] designed is classifying cases brought before the ECtHR to: (a) $D_{\overline{\mathcal{N}}}$: cases that have *violated* an article of the Convention; (b) $D_{\mathcal{N}}$: cases that have *not* violated that article. For instance, during the training phase of their algorithm, they "*fed*" the algorithm with past cases of violations of *Article 3*:

> **Article 3 (Prohibition of torture)**: "*No one shall be subjected to torture or to inhuman or degrading treatment or punishment.*"

By doing so, the algorithm identified words that appear with high frequency in those cases (e.g., "*injury*", "*ukraine*", "*detainee*", "*food*"). I.e., the patterns the algorithm was using to distinguish the two categories "*violation*" and "*no violation*" were *patterns of words*. Afterwards, during the testing phase, they "*fed*" the algorithm with new cases of alleged violations of *Article 3*. Whenever the algorithm encountered one of the aforementioned words in a new case, it was raising the probability of that case being judged by the ECtHR as a violation of *Article 3*. Consequently, when the algorithm was classifying a new case d as a violation of *Article 3*, the explanation was the *particular fact* f ="*The words* w_1, w_2, ..., w_n *appear in* d *and they also appear with high frequency in past cases violating Article 3.*".

Now facts of the form of f are facts about *Article 3* itself. Since the term "*Article 3*" is mentioned in f, for f (a *particular fact*) to stand in a subsumptive relation with *Article 3* (a *norm*), "*Article 3*" should either be a term appearing in *Article 3* itself (*self-reference*) or it should belong to a concept mentioned in *Article 3*. Neither of the two is the case. In other words, unless a norm is self-referential - either by including the norm itself as a term or by including a concept to which the norm belongs to - the possibility of a fact of the form of f standing in a subsumptive relation with that norm is excluded.[11] Hence, such a fact can not be part of the propositions of the hierarchy $H_{P,2}$ since those propositions stand in subsumptive-deductive derivation relations with norms of $H_{P,2}$. Consequently, an explanation that includes facts like f is *not* an explanation that adheres to the normative model EXP_{BC} *at all* - not even in *degree 1*.

3.2.2. EXP_{BC} and rule-based programming

In contrast with ML algorithms, *ruled-based models* can provide explanations of *degree 3*. A rule-based model is a model that consists of *rules* $\phi(x) \Rightarrow \psi(x)$ and *facts* $\chi(a)$, where ϕ, ψ, χ are propositional functions,[12] x is a variable and a a term without free variables. Whenever $\phi(a)$ is true, then for every rule i of the form $\phi(x) \Rightarrow \psi_i(x)$ we have that $\psi_i(a)$ is true [15].

When they appear in the code of a programme Π, *rules* can be interpreted as *norms* and *facts* as *propositions* [15]. Specifically, $\phi(x) \Rightarrow \psi(x)$ can be interpreted as "*Whenever* $\phi(x)$ *is the case, then* $\psi(x)$ ***should be*** *the case.*" and the fact $\phi(a)$ can be interpreted as "$\phi(a)$ ***is*** *the case.*". Usually, the *output* has *only* facts like $\psi(a)$. When in the output, facts can also be interpreted as *norms*: "*According to the programme* Π, $\psi(a)$ ***must be*** *the case.*". This flexibility in the interpretation of the output is on par with the remark about being able to construe a judgement as both a norm and a proposition mentioned at ¶3 of (SUM.3)'s application to legal practice in §2.3.

---

[11] The generalised version of f for all classification ML algorithms is: f ="*The* ***patterns*** p_1, p_2, ..., p_n *appear in data* d *and they also appear with high frequency in past cases of category* C*. Hence,* d *belongs to* C*.*"

[12] Their arities can vary.

Let's see now a formalisation of Example 1 using rule-based programming. Assume that $\psi(x)$:="*x pays taxes in Italy on their worldwide income.*", $\phi(x)$:="x *lives in Italy for more than 183 consecutive days over a 12-month period.*". Then, we can formalize Example 1 as follows:

$$\begin{array}{l} \text{line 1. } \phi(x) \Rightarrow \psi(x) \\ \text{line 2. } \phi(Alice) \\ \hline \text{output: } \{\psi(Alice)\} \end{array}$$

Clearly, this formalisation provides an explanation of *degree 3*: it contains the norms (*line 1*, *output*), propositions (*line 2*), and the derivation relation among them.

4. ASP causal explanations supervening on EXP_{BC}

4.1. ASP in a nutshell

ASP is a rule-based programming method that uses first-order logic (FOL) language. As such, an ASP programme Π is a set of rules of the form $body \Rightarrow head$, where $head$ is an atom and the body consists of combinations of literals L_i, where each L_i can either be an atom a or its *default* negation $\sim a$. An atom is said to be *proven* whenever it appears in the head of rule whose body is satisfied. When we have the edge case where the body of a rule is empty (i.e., $\Rightarrow head$), the head is considered proven under *any* circumstances. Hence, we call it a *fact* [20]. The notion of a proven atom is central for ASP since the output of any ASP programme Π is logical models that include *only* those atoms which have been proven. Those models are called *answer sets* and hence the name "*answer set programming*". Finally, as a FOL programming method, ASP syntax includes *relations* (e.g., R) and *functions* (e.g., f) of arities n and n' respectively symbolised as R/n and f/n'. For instance, a common way of representing *graphs* in ASP is to employ the following two relations: (i) a predicate $node/1$ which singles out the atoms which are nodes; (ii) a binary relation $edge/2$ which signifies the existence of a directed edge from its first to its second argument (e.g., $edge(a, b)$ represents $a \rightarrow b$). For a more detailed & quick introduction to the basics of ASP *see* [20].

Let's see now how the rule-based model of Example 1 proposed in §3.2.2 can be realised *via* ASP. Assume that the predicate $italy183/1$ stands for ϕ, the predicate $taxes2Italy/1$ stands for ψ. Then the programme $\Pi^{ex_1} = \{italy183(X) \Rightarrow taxes2Italy(X), italy183(alice)\}$ has only one answer set: $\mathcal{AS}^{ex_1} = \{taxes2italy(alice), italy183(alice)\}$. Specifically, since $italy183(alice)$ is a fact it has to belong to every answer set of Π^{ex_1}. At the same time, since the body of the rule $italy183(X) \Rightarrow taxes2Italy(X)$ is satisfied for $X = alice$, then its head $taxes2Italy(X)|_{X=alice}$ must also belong to every answer set of Π^{ex_1}. Since there are no more *proven* atoms, the only answer set compatible with Π^{ex_1} ends up being that of $\mathcal{AS}^{ex_1}$.

4.2. Causal models of causal structures supervening on $H_{P,2}$

An argument against CMS_{LP} could be that in the actual legal practice the derivation relations among the propositions and norms of $H_{P,2}$ are *not* exhausted in subsumptive-deductive inference. Indeed, by looking in any book on legal reasoning (*see* e.g., [21], [22]) one can see

that there is a plurality of reasoning methods used in the actual practice: analogical/evidential reasoning, deontic logic, counterfactuals, etc. However, CMS_{LP} is *not* a normative model of how legal practitioners should *reason* in their practice (e.g., by using abduction [22]). Instead, it is a normative model on how legal practitioners should *explain* the outcome of their practice. I.e., it is a *meta*-analysis of how they actually reason to conclude to that outcome.

Having said that, one can still *reconstruct* the proposed normative explanations to reflect a reasoning method different than subsumption-deduction in a way that it is *still* clear which is the hierarchical relation $H_{P,2}$ among the propositions and norms involved in that reconstruction. In what follows, I will do so for the case of causal inference using ASP. Note that causal inference is of prime importance for explanations of legal judgements. The most characteristic such explanation is the alleged causal relation between the defendants' actions (alleged *cause*) to the applicants' alleged harm (alleged *effect*).

Before reconstructing the proposed subsumptive-deductive explanations to causal explanations, I need to decide on the the definition of causal inference; I have to know *what* I model before modelling it. There is a diverse plethora of available definitions motivated by different *metaphysical* conceptualisation of causation. Considering that, I am choosing a metaphysically *neutral* definition so as to be compatible with many such conceptualisations.

Definition 4.1 (Cause). X is a *cause* of Y iff: (i) it is possible to intervene on X; (ii) under some such possible intervention on X, changes in the value of X are associated changes in the value of Y. [23, p.3583]

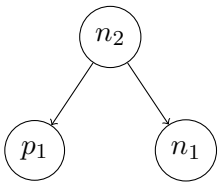
Since the *desideratum* is a conceptualisation of causal inference to reflect the hierarchy $H_{P,2}$ the variables X and Y of Definition 4.1 will be propositions and norms that are part of that hierarchy. More precisely, considering clauses (i) and (ii) of Definition 4.1, a proposition/norm X will be the cause of another proposition/norm Y *iff*

4.1.i′ *it is possible intervene to the value of* X: The values of X and Y in this situation are *truth values* - those of truth_2. Moreover, "*possibility*" in this context is a *conceptual* possibility; it does not reflect the *actual* state of affairs, but it is a *counterfactual* case of another *non*-actual state of affairs. For instance, it may be the case that in the *actuality* Alice has lived in Italy for more than 183 consecutive days over a 12-month period, but it is also possible to *conceptualise* a counterfactual situation in which Alice left on day 182 for a conference on logical programming in Haifa, Israel.

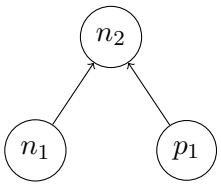
4.1.ii′ *by intervening in the value of* X*, we intervene in the value of* Y: This is the point where the hierarchy $H_{P,2}$ comes into play. $H_{P,2}$ is induced by *derivation* relations: norms situated higher are *derived* from norms and propositions situated lower *via* subsumption-deduction. *See* for instance the left graph of Figure 4.1. The direction of the edges $n_2 \rightarrow n_1$ and $n_2 \rightarrow p_1$ exhibits that n_2 is *derived from* (or *grounded on*) n_1 and p_1. Since n_2 is "*derived*" it is the *true* conclusion of an argument whose premisses are also *true*. In other words, all the variables in the current state of affairs (that of $H_{P,2}$) have the value "*true*". Hence, the *only* intervention to any such variable that we can make is that of setting it *false*. Consequently, X will *cause* Y *iff* whenever we set its truth value to *false* we have that Y's truth value also become *false*.

The foregoing conceptualization of causation reflects the hierarchical structure $H_{P,2}$ since X *being the cause of* Y implies that X: (i) is situated lower in the hierarchy than Y; (ii) is one of the premisses from which Y is derived *via* subsumption-deduction.

Let's proceed with modelling this conceptualisation of causality. Assume a *causal structure* $\mathfrak{C}$ that exists in the *actual* world. As a causal structure I define a collection of *causal relata* and the *causal relations* among them. From 4.1.ii′, we know that such a structure *supervenes* on the hierarchy $H_{P,2}$ and that the causal relata are the involved propositions/norms. A *causal graph* $\mathcal{G}(\mathfrak{C}) = \langle N(\mathfrak{C}), E(\mathfrak{C})\rangle$ is a graph whose nodes $N(\mathfrak{C})$ are representations of $\mathfrak{C}$'s causal relata and whose edges $E(\mathfrak{C})$ are representations of the *direct* causal relation[13] among those relata and hence, it can serve as a *model* of $\mathfrak{C}$.



$H_{P,2}$ of Example 1.
$\mathfrak{C}$ supervenes on $H_{P,2}$



$\mathcal{G}(\mathfrak{C})$

Figure 4.1.

A prominent method of constructing $\mathcal{G}(\mathfrak{C})$ is to bookeep a *list* $L_{\mathfrak{C}}$ of all the independencies between the *causal relata* of $\mathfrak{C}$ and then, construct $\mathcal{G}(\mathfrak{C})$ in such a way that expresses those and *only* those independencies. I.e., assuming that we also have a list $L_{\mathcal{G}(\mathfrak{C})}$ of the independencies among $\mathcal{G}(\mathfrak{C})$'s nodes, the *ideal* is that for every set of independent *relata* in $\mathfrak{C}$ their representations in $\mathcal{G}(\mathfrak{C})$ are also independent and *vice versa.* That requirement presupposes at least two distinct *formal* notions of independence - one for $\mathfrak{C}$'s causal relata and one for $N(\mathfrak{C})$ - and that both notions *explicate* formally the *same* concept of independence.

Let's start with the concept of independence we want to formalise. Assume three disjoint sets of causal relata X, Y, Z. Then, an independency is a proposition of the form: "*Knowing* Z *renders* X *irrelevant to* Y*.*" [25, p.5]. We symbolise this proposition as $I_X^{\mathfrak{C}} Y|Z$. Similarly, for three disjoint sets of nodes X', Y', Z' an independency is a proposition of the form "*Knowing the value of the variables in* Z' *renders the values of the variables in* X' *irrelevant to the values of the variables in* Y'*.*" symbolised as $I_{X'}^{\mathcal{G}(\mathfrak{C})} Y'|Z'$. Verma & Pearl [26, p.354] provide a more "*formal*" definition which can be adjusted to the context of this paper as follows:

Definition 4.2 (Independency $\mathbf{I_X Y|Z}$). Assume an *ordered* set of variables V and three disjoint subsets of V $X = \{x_1, x_2, ..., x_n\}, Y = \{y_1, y_2, ..., y_k\}, Z = \{z_1, z_2, ..., z_m\}$ such that all variables of $X \cup Z$ are situated lower than Y's variables in that ordering. Then, Y is *independent* of X given Z *iff* when *assigning* specific values to Z's variables ($\{z_1 = Z_1, z_2 = Z_2, ..., z_m = Z_m\}$) the value Y_j for every variable of $y_j \in Y$ will be the same for any set of values $\{X_1, X_2, ..., X_n\}$ we assign to the variables of X.

[13] An informal definition of "*direct cause*" is the following: C is a *direct* cause of E *iff* there is no causal relatum that mediates between C and E [24, p.20]. A paradigmatic example of that is the following: billiard ball b_1 hits billiard ball b_2 which hits billiard ball b_3. b_1 *causes* b_3 to move, but it is *not* a direct cause since there is another causal relatum - that of b_2 - that mediates between them.

Although this definition is more "*formal*", it is still not complete since it leaves many open questions - which are not in the scope of this paper to be answered - like what kind of orderings of the set V are acceptable. Despite that, it is still quite an insightful definition since it reveals two important prerequisites for explicating formally the concept of *independence*:

(I.1) there is a *functional dependence* of Y to both X and Z: $Y = f(X, Z)$
(I.2) there needs to be an *ordering* of the variables such that the variables situated higher in that ordering can *potentially* be functionally depended in the variables situated lower in that ordering.

From 4.1.ii′, we infer that the ordering of (I.2) is that of the hierarchy $H_{P,2}$ and that the functional dependence of (I.1) is that of a truth assignment function. In the literature of causal modelling, $I_X^{\mathfrak{C}}Y|Z$ is usually *probabilistic* independence and the functional dependence $Y = f(X, Z)$ is *asymmetric*. I.e., although Y is functionally dependent on both X and Z, the inverse does not hold. The latter needs to be a requirement for our model as well since it represents an inherent asymmetric property of causal *relata*: the effect *depends* on the cause but not the inverse [27].

4.3. ASP modelling of causal graphs & $\mathbf{EXP_{BC}}$

The construction of causal graphs using ASP[14] is performed in the following three steps:

(CM.1) the existence of a direct edge between two nodes $n, m \in N(\mathfrak{C})$ is encoded as an atom $edge(n, m)$
(CM.2) we place restrictions on those atoms to force them respect the independencies $L_{\mathcal{G}(\mathfrak{C})}$
(CM.3) an answer set solver returns all stable models that satisfy those restrictions. The collection of edges in each such model is the required causal graph.

The ASP programme described in those three steps does not output a binary classification, but causal graphs. However, with a few extra steps we can induce from it a binary classification algorithm. Assume for instance that we want to use ASP to perform the same binary classification as in the example used in §3.2.1: the *input* is the facts of a case and the *output* is whether or not there has been been a violation of the Convention's Article 3. Firstly, we construct an ASP programme $\Pi^{G(\mathfrak{C})}$ that outputs causal graphs based on the process described above. For that, we use past cases of violations of Article 3 to identify the independencies $I_X^{\mathfrak{C}}Y|Z$. Let's postpone the discussion on the process of identifying those independencies for the conclusion of the paper. Now from 4.1.ii′ and Definition 4.2, we can infer that an independency of the form $I_X^{\mathcal{G}(\mathfrak{C})}Y|Z$, means that if values of X's variables change to false the values of all Y's variables will still remain true as long as the values of variables of Z remain true. Having that in mind, we can construct a binary classification ASP programme Π^{BC} in the following way:

$$\Pi^{BC} = \Pi^{G(\mathfrak{C}),2} \cup \Pi^{facts} \cup \Pi^{R}$$

- where $\Pi^{G(\mathfrak{C}),2}$ is an ASP programme that consists of the edges outputted by $\Pi^{G(\mathfrak{C})}$
- Π^{facts} is an ASP coding of the facts of a case

[14] *See* e.g. [28, 29, 30] and [31, §7.2.1].

- Π^R is a set of rules which given an independence $I_X Y|Z$ requires the variables Y to be in the answer set in case that the variables of Z are in Π^{facts}. A naive such set of rules could be that if all direct causes of a are true then a has to be true as well. The required output - the violation of Article 3 - is encoded in Π^R as an atom. In case that it appears in the answer sets, then Π^{BC} has outputted that we have a violation of Article 3.

It is part of my Thesis - an ongoing project - to come up with rules for Π^R and eventually construct an example of Π^{BC} regarding hate speech cases of the ECtHR.

5. Conclusion

What I would like to remark in the conclusion is that in terms of explainability the most difficult black box to whiten in the foregoing proposal of causal modelling is the process of *deciding the independencies*. The common practice is to consider $I^{\mathfrak{C}}$ to be a *probabilistic* independence. However, that leaves room for the same kind of criticism than the one against machine learning I used in §3.2.1; $I^{\mathfrak{C}}$ is based on frequentistic arguments of the form of statistical independence tests: "*In n out of the m examined cases, Y was independent of X when we condition on Z and according to the statistical test S they should be conditionally independent.*".

In order to overcome this problem we are left without any other choice than inquiring for different forms of functional (in)dependencies than the probabilistic ones. Conveniently, a plurality of important theorems that allow probabilistic independencies to be translated successfully to graph-theoretical independencies have been proven for functional dependencies that satisfy certain properties (e.g., *symmetry* $I_X Y|Z \leftrightarrow I_Y X|Z$ and *decomposition* $I_{X\cup W} Y|Z \rightarrow I_X Y|Z$ [32]) and not for probability functions in particular. Hence, a good point to start is to find functions that still satisfy those properties and whose semantic interpretation is compatible with EXP_{BC}. Since the variables of $G(\mathfrak{C})$ are norms and propositions, it is rather intuitive that *truth functions* - not *per se* those of classical logic - have such semantic interpretations.

Acknowledgments

Thanks to the guidance of Arianna Betti, Ronald de Haan , and Katrin Schulz as well as the insightful feedbacks of Derek So (Wing Yi) and of the reviewers of the ASPOCP 2022 Workshop.

References

[1] European Commission for the Efficiency of Justice, European ethical Charter on the use of artificial intelligence in judicial systems and their environment, printed by the Council of Europe, 2019.

[2] D. Remus, F. S. Levy, Can robots be lawyers? computers, lawyers, and the practice of law, Georgetown journal of legal ethics 30 (2016) 501+.

[3] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, Predicting judicial decisions of the european court of human rights: a natural language processing perspective, PeerJ Comput. Sci. 2 (2016) e93.

[4] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the european court of human rights, Artificial Intelligence and Law 28 (2020) 237–266. doi:doi.org/10.1007/s10506-019-09255-y.
[5] M. C. C. F. V. A. L. S. Masías VH, Valle M, Modeling verdict outcomes using social network measures: The watergate and caviar network cases., PLoS ONE (2016) e0147248. doi:10.1371/journal.pone.0147248.
[6] Lawformer: A pre-trained language model for chinese legal long documents, AI Open 2 (2021) 79–84. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000176. doi:10.1016/j.aiopen.2021.06.003.
[7] G. P. H.-F. N. K. R. Z. J. . Z. H. Chhatwal, R. (2018).
[8] F. Wei, H. Qin, S. Ye, H. Zhao, Empirical study of deep learning for text classification in legal document review, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3317–3320. doi:10.1109/BigData.2018.8622157.
[9] G. Sartor, A. Loreggia, The impact of algorithms for online content filtering or moderation - upload filters, Study requested by the JURI Committee. European Parliament Think Tank, 2020. URL: https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101.
[10] B. Adrien, M. Lognoul, A. de Streel, B. Frénay, Legal requirements on explainability in machine learning, Artificial Intelligence and Law 29 (2021) 149–169. doi:10.1007/s10506-020-09270-4.
[11] Álvaro Núñez Vaquero, Five models of legal science, Revus 19 (2013). doi:https://doi.org/10.4000/revus.2449.
[12] W. R. de Jong, A. Betti, The classical model of science: a millennia-old model of scientific rationality, Synthese 174 (2010) 185–203. doi:10.1007/s11229-008-9417-4.
[13] L. Alexander, E. Sherwin, Demystifying legal reasoning, Cambridge University Press, 2008.
[14] N. MacCormick, Legal deduction, legal predicates and expert systems, International Journal for the Semiotics of Law 5 (1992) 181–202. doi:10.1007/BF01101868.
[15] G. Governatori, A. Rotolo, G. Sartor, Logic and the law: philosophical foundations, deontics, and defeasible reasoning, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden, L. van der Torre (Eds.), Handbook of deontic logic and normative systems, volume 2, College Publications, 2021.
[16] C. E. Alchourrón, Limits of logic and legal reasoning, in: C. Bernal, C. Huerta (Eds.), Essays in legal philosophy, Oxford University Press, [1992] 2015. doi:10.1093/acprof:oso/9780198729365.003.0017.
[17] R. Dworkin, Justice for hedgehogs, Belknap Press of Harvard University Press, 2011.
[18] F. Schroeter, L. Schroeter, K. Toh, A new interpretivist metasemantics for fundamental legal disagreements, Legal Theory 26 (2020) 62–99. doi:10.1017/S1352325220000063.
[19] C. M. Bishop, Pattern recognition and machine learning (Information, science and statistics), Springer-Verlag, 2006.
[20] M. Gebser, R. Kaminski, B. Kaufmann, T. Schaub, Answer set solving in practice (2012). doi:10.2200/S00457ED1V01Y201211AIM019.
[21] G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini, D. Walton (Eds.), Handbook of legal reasoning and argumentation, Springer, Dordrecht, 2018.
[22] D. Walton, Legal argumentation and evidence, The Pennsylvania State University Press,

2002.
[23] J. Woodward, Methodology, ontology, and interventionism, Synthese 192 (2015) 3577–3599. doi:10.1007/s11229-014-0479-1.
[24] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, 2 ed., MIT Press, 2000.
[25] J. Pearl, A. Paz, GRAPHOIDS: A Graph-based Logic for Reasoning about Relevance Relations, University of California computer science department, technical Report 850038 (R-53), 1985.
[26] T. Verma, J. Pearl, Causal networks: semantics and expressiveness, in: Proceedings of the fourth workshop on uncertainty in artificial intelligence, 1988, pp. 352–359. doi:10.48550/ARXIV.1304.2379.
[27] N. Cartwright, Against modularity, the causal markov condition, and any link between the two: comments on Hausman and Woodward, British Journal for the Philosophy of Science 53 (2002) 411–453. doi:10.1093/bjps/53.3.411.
[28] Z. Zhalama, J. Zhang, F. Eberhardt, W. Mayer, M. J. Li, Asp-based discovery of semi-markovian causal models under weaker assumptions, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19, AAAI Press, 2019, p. 1488–1494. doi:10.5555/3367243.3367245.
[29] A. Hyttinen, F. Eberhardt, M. Järvisalo, Constraint-based causal discovery: conflict resolution with answer set programming, in: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI'14, AUAI Press, Arlington, Virginia, USA, 2014, p. 340–349. doi:10.5555/3020751.3020787.
[30] S. Triantafillou, I. Tsamardinos, I. G. Tollis, Learning causal structure from overlapping variable sets, in: In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 2010, p. 860–867.
[31] J. Peters, D. Janzing, B. Schölkopf, Elements of causal inference: foundations and learning algorithms, The MIT Press, 2017.
[32] J. Pearl, T. S. Verma, The logic of representing dependencies, in: Proceedings of the 6th national conference on A.I. (AAAI-87), volume 1, 1987, pp. 374–379.