AI4J – Artificial Intelligence for Justice

August 30, 2016

The Hague, The Netherlands

Workshop at the 22nd European Conference on Artificial Intelligence (ECAI 2016)

Workshop chairs

Floris Bex
Department of Information and Computing Sciences, Utrecht University

Tom van Engers
Leibniz Center for Law, Faculty of Law, University of Amsterdam

Henry Prakken
Faculty of Law, University of Groningen;
Department of Information and Computing Sciences, Utrecht University

Bart Verheij
Institute of Artificial Intelligence and Cognitive Engineering,
University of Groningen

Artificial intelligence is currently in the centre of attention of legal professionals. An abundance of startup companies explore the application of AI techniques in the domain of law, and there is even talk of artificially intelligent legal assistants disrupting the legal market space. Factors driving the increased attention for legal AI include:

- Technological breakthroughs in machine learning, natural language processing, ubiquitous computing, data science, and argumentation technology;
- The changing attitude towards technology in the legal domain;
- The much increased availability of legal data on the internet;
- The recent success of AI applications in the private and public domain;
- The success of technology supporting access to law, legal empowerment, and transparency;
- The increased need for norms embedded in technology (autonomous driving and warfare, big data analysis for crime fighting and counterterrorism).

The aim of this workshop is to investigate opportunities and challenges in AI applied to the law, with a particular focus on the relevance of the recent technological breakthroughs for AI & Law research and for legal practice. Questions addressed include the following:

- How can AI & Law research contribute to improving legal work in, for example, courts, law firms, public administration, police practice and businesses?
- How should AI & Law research change in light of the recent research breakthroughs and technological developments? For example, how can traditional research on legal knowledge bases, legal reasoning and legal argument be combined with data science, machine learning and natural language processing?

The law has been a longstanding application field for AI research. The biennial International conferences on AI & Law (ICAIL) started in 1987, the annual JURIX conferences on Legal Knowledge and Information Systems in 1988, and the journal Artificial Intelligence and Law was founded in 1992. Many ideas that are currently being commercially developed were first explored by AI & Law researchers, such as legal information retrieval, statistical analysis of legal data, automated contract drafting, automated processing of permit or welfare applications in public administration, and systems for regulatory compliance.

Some particular topics of relevance to the focus of the workshop are:

- Open data, linked data, big data;
- e-Discovery;
- Legal argument mining;
- Automated contract drafting;
- Computable contracts;
- Decision support for legal, forensic and police practice;
- Computational law.

Accepted papers (full, position and short)

Invited speaker

Karl Branting
AI & Law from a Data-Centric Perspective

Data-centric techniques are applicable both to extracting information latent in legal document collections and to finessing some of the central challenges confronting logical models, including the practical difficulties of formalizing legal rules in logic at scale and the mismatch between legal predicates and ordinary parlance. This talk surveys data-centric techniques and applications, distinguishes legal tasks amenable to these techniques from those requiring logic-based analysis, and describes the rapidly growing commercial interest in these techniques.

Program committee

Kevin Ashley, University of Pittsburgh, USA
Katie Atkinson, University of Liverpool, UK
Trevor Bench-Capon, University of Liverpool, UK
Karl Branting, The MITRE Corporation, USA
Pompeu Casanovas, Universitat Autnoma de Barcelona, Spain; Deakin University, Australia
Jack G. Conrad, Thomson Reuters, USA
Enrico Francesconi, ITTIG-CNR, Italy
Tom Gordon, Fraunhofer FOKUS, Germany
Guido Governatori, NICTA, Australia
Matthias Grabmair, Intelligent Systems Program, University of Pittsburgh, USA
Jeroen Keppens, Kings College London, UK
David Lewis, Chicago, USA
Monica Palmirani, CIRSFID, Italy
Dory Reiling, Court of Amsterdam, The Netherlands
Erich Schweighofer, University of Vienna, Austria
Jaap van den Herik, Leiden University, The Netherlands
Serena Villata, INRIA Sophia Antipolis, France
Radboud Winkels, University of Amsterdam, The Netherlands
Adam Wyner, University of Aberdeen, UK

Financial support

International Association for Artificial Intelligence and Law (IAAIL)
JURIX Foundation for Legal Knowledge Based Systems (JURIX)
BNVKI BeNeLux Vereniging voor Kunstmatige Intelligentie (BNVKI)
ALICE Institute for Artificial Intelligence and Cognitive Engineering (ALICE)

Toward Machine-Understandable Contracts

Sudhir Agarwal[1] and **Kevin Xu** [2] and **John Moghtader** [3]

Abstract. We present Contract Definition Langauge, a novel approach for defining contracts declaratively in a machine-understandable way to achieve better comprehensibility and higher efficiency of reasoning, analysis and execution of contracts through higher degree of automation and interoperability. The effect of representing contracts with our Contract Definition Language is not only a significant reduction of legal transaction costs, but it also opens a variety of new options to create better contracts. As a proof of concept, we also present our modeling of two US statutes (FERPA/SOPIPA/COPAA and HIPAA) as well as our prototypes for validity checking and hypothetical analysis of contracts according to those statues.

1 Introduction

The conventional view is that the automation of contract creation, execution, and compliance is beyond the capabilities of today's technologies. This view stems from a longstanding tradition of contracting practices, where all terms and conditions are expressed and interpreted in natural language, often in obtuse, inaccessible legalese that can only be deciphered by lawyers and judges. Many legal scholars have called for more usable, interactive tools to make better sense of contracts [1, 4, 8, 11, 14]. Contract scholars have defined a need for user-facing tools that would make contracts more understandable and actionable [4, 12], as well as a more modular and machine-readable approach to future contracting practices [3, 15, 16].

The impetus for the Computable Contracts research at Stanford [4], and Computational Law [5] more broadly, is a vision where computers, or people with the help of computers, are able to rapidly understand the implications of contracts in their legal context, in order to make optimal decisions accordingly, on a potentially large and complex scale. This vision, if realized, will dramatically improve access to justice in the legal system.

In our definition, computable contracts have the following main features:

1. **Machine-understandable**: In contrast to traditional natural language contracts, computable contracts have logic-based formal semantics which enables use of machines to automatically reason over contracts.
2. **Declarative**: In contract to hard-coding contract terms with a procedural programming language such as Java and C++, computable contracts are defined declaratively. Thus computable contracts can be better comprehensible by legal professionals as well as the clients as the declarative nature is closer to the way domain knowledge is specified.
3. **Executable**: Computable contracts are executable like procedural code. Thus, computable contracts do not need to be translated to programmed in a traditional programming language which reduces costs and errors as there is not need to manage two separate versions (one for humans, another for machines) of the same contract.
4. **Interoperable**: Computable contracts are interoperable in the sense that they use shared vocabulary for referring to real world objects, thus enabling automating reasoning over multiple different contracts that may have interdependencies or even conflicting terms.

In this paper, we present an approach for formulating, analyzing, and executing contracts more efficiently and effectively by enabling a high degree of automation. We introduce a Logic Programming based Contract Definition Language (CDL). CDL makes possible automated reasoning over legal substance. We have identified the following four types of reasoning tasks with contracts:

1. **Validity Check**: Validity checking determines whether a contract satisfies all the constraints it must satisfy. For example, a marriage contract between an adult and a minor is invalid in most countries because the law of the countries does not allow such contracts. Constraints that need to be satisfied can be defined with CDL as well.
2. **Hypothetical Analysis**: In many cases, a user (e.g. a legal professional, a client, a customer, an employee etc.) wishes to understand a given contract or a set of contracts for a situation. For example, an employee who is not keeping very good health may be interested in knowing what happens when he/she has to take more sick leave than mentioned in the contract. Hypothetical analysis roughly provides an answer to the question: What are the implications (obligations/rights) of laws and/or contract terms in a particular, given or hypothetical, situation?
3. **Utility Computation**: Terms in contracts often have a different utility for the involved parties. The utility depends on the party's preferences. When a party's preferences are known, the utility of a contract for the party can be computed automatically.
4. **Planning**: A contract can be seen as a set of constraints on involved parties' behavior. When the goal of a party is known, a plan, i.e. a sequence of actions, can be computed automatically such that the execution of the sequence of actions would lead the party to its desired goal state. Planning problem has been extensively studied as a discipline of Computer Science and Artificial Intelligence [5], and it might be possible to adopt one of the existing techniques for our purpose.

As far as reasoning with contracts is concerned, in this paper, we focus on validity check and hypothetical analysis only. They allow

[1] CS Dept., Stanford University, USA. Email: sudhir@cs.stanford.edu
[2] Law School, Stanford University, USA. Email: kevin.s.xu@gmail.com
[3] Law School, Stanford University, USA. Email: jmoghtader@gmail.com
[4] `http://compk.stanford.edu`
[5] `http://complaw.stanford.edu`

automated support while still leaving the decision-making and implementation to the user, and are also required by other reasoning tasks.

The paper is organized as follows. We first give an overview of foundations upon which our technique is built. Then, we present the syntax and semantics of our Contract Definition Language (CDL). We present two case studies involving modeling with CDL and automatically reasoning about two U.S. Federal statutes, FERPA and HIPAA. We conclude and identify next steps after a discussion of related work.

2 Foundations

In this section we give a short overview of the syntax and intuitive semantics of deductive databases and logic programs, two foundational techniques upon which we build our Contract Definition Language.

2.1 Databases

The *vocabulary* of a database is a collection of object constants, function constants, and relation constants. Each function constant and relation constant has an associated arity, i.e. the number of objects involved in any instance of the corresponding function or relation. A term is either a symbol or a functional term. A functional term is an expression consisting of an n-ary function constant and n terms. In what follows, we write functional terms in traditional mathematical notation - the function followed by its arguments enclosed in parentheses and separated by commas. For example, if `f` is a binary function constant and if `a` and `b` are object constants, then `f(a,a)` and `f(a,b)` and `f(b,a)` and `f(b,b)` are all functional terms. Functional terms can be nested within other functional terms. For example, if `f(a,b)` is a functional term, then so is `f(f(a,b),b)`. A datum is an expression formed from an n-ary relation constant and n terms. We write data in mathematical notation. For example, we might write `parent(art,bob)` to express the fact that Art is the parent of Bob. A dataset is any set of data that can be formed from the vocabulary of a database. Intuitively, we can think of the data in a dataset as the facts that we believe to be true in the world; data that are not in the dataset are assumed to be false.

2.2 Logic Programs

The language of logic programs includes the language of databases but provides additional expressive features. One key difference is the inclusion of a new type of symbol, called a *variable*. Variables allow us to state relationships among objects without explicitly naming those objects. In what follows, we use individual capital letters as variables, e.g. `X`, `Y`, `Z`. In the context of logic programs, a term is defined as an object constant, a variable, or a functional term, i.e. an expression consisting of an n-ary function constant and n simpler terms. An atom in a logic program is analogous to a datum in a database except that the constituent terms may include variables. A *literal* is either an atom or a negation of an atom (i.e. an expression stating that the atom is false). A simple atom is called a *positive* literal, The negation of an atom is called a *negative* literal. In what follows, we write negative literals using the negation sign `~`. For example, if `p(a,b)` is an atom, then `~p(a,b)` denotes the negation of this atom. A *rule* is an expression consisting of a distinguished atom, called the head and a conjunction of zero or more literals, called the *body*. The literals in the body are called *subgoals*. In what follows, we write rules as in the example `r(X,Y) :- p(X,Y) & ~q(Y)`. Here, `r(X,Y)` is the head, `p(X,Y) & ~q(Y)` is the body; and `p(X,Y)` and `~q(Y)` are subgoals.

Semantically, a rule states that the conclusion of the rule is true whenever the conditions are true. For example, the rule above states that `r` is true of any object `X` and any object `Y` if p is true of `X` and `Y` and `q` is not true of `Y`. For example, if we know `p(a,b)` and we know that `q(b)` is false, then, using this rule, we can conclude that `r(a,b)` must be true.

3 Contract Definition Language (CDL)

CDL descriptions are open logic programs. While a traditional logic program is typically used to specify views and constraints on a single database state, CDL descriptions can specify a state-transition system. CDL is expressive enough to define a Turing machine. The declarative syntax and formal semantics of CDL makes CDL descriptions easier to comprehend and maintain. The executability of CDL descriptions makes it superfluous to hard-code contract terms with procedural code as well as makes CDL a promising alternative for defining self-executable contracts such as Ethereum Smart Contracts [6, 10].

The basis for CDL is a conceptualization of contracts in terms of entities, actions, propositions, and parties. Entities include objects relevant to the state of a contract are usually represented by object constants in CDL. In some cases, we use compound terms to refer to entities. Actions are performed by the parties involved in the contract. As with entities, we use object constants or compound terms to refer to primitive actions. Some actions may not be legal in every state. Propositions are conditions that are either true or false in each state of a contract. In CDL, we designate propositions using object constants or compound terms. Parties are the active entities in contracts. Note that, in each state, some of the contract's propositions can be true while others can be false. As actions are performed, some propositions become true and others become false. On each time step, each party has a set of legal actions it can perform and executes some action in this set. In CDL, we usually use object constants (in rare cases compound terms) to refer to parties. In CDL, the meaning of some words in the language is fixed for all contracts (the contract-independent vocabulary) while the meanings of all other words can change from one contract to another (the contract-specific vocabulary).

There are the following contract-independent structural relation constants.

1. `role(r)` means that r is a role in the contract.
2. `base(p)` means that `p` is a base proposition in the contract.
3. `percept(r,p)` means that `p` is a percept for role `r`.
4. `input(r,a)` means that `a` is an action for role `r`.

To these basic structural relations, we add the following relations for talking about steps.

1. `step(s)` means that `s` is a step.
2. `successor(s1,s2)` means that step `s1` comes immediately before step `s2`.
3. `true(p,s)` means that the proposition `p` is true on step `s`.
4. `sees(r,p,s)` means that role `r` sees percept `p` on step `s`.
5. `does(r,a,s)` means that role `r` performs action `a` on step `s`.
6. `legal(r,a,s)` means it is legal for role `r` to play action `a` on step `s`.

7. `goal(r,n,s)` means that player has utility `n` for player `r` on step `s`.
8. `terminal(s)` means that the state on step `s` is terminal.

The truth of propositions in the initial state can be stated using true with the first step as the step argument; and update rules can be stated using `true` and `successor`. Just how this works should become clear from the following modeling of a part of the well known board game Tic-Tac-Toe.

There are two roles say `black` and `white` who make their moves alternatively.

```
role(white)
role(black)
true(control(black),N) :-
    true(control(white),M) &
    successor(M,N)
true(control(black),N) :-
    true(control(white),M) &
    successor(M,N)
```

If `white` marks a cell in a state, then that cell has a `x` in the next state. Analogously, if `black` marks a cell in a state, then that cell has a `o` in the next state. Further rules not shown below ensure that all other cells carry their previous markings to the next state.

```
true(cell(I,J,x),N) :-
    does(white,mark(I,J),M) &
    successor(M,N)
true(cell(I,J,o),N) :-
    does(black,mark(I,J),M) &
    successor(M,N)
```

CDL interprets negations as failure and does not allow negations or disjunctions in the head. While negation as failure could be a limitation in some scenarios, in many scenarios one can safely make a closed-world assumption. The inability to express disjunctions in the head can be a limitation in some cases, but in many cases, the regulations are laws and regulations are definite. In many cases, once can introduce a new atom to represent the union of the disjunts. Exceptions can be modeled with CDL indirectly by introducing an auxiliary view and adding its negation in the body of the appropriate rules.

4 Case Studies

Thus far, we have modeled two sets of U.S. Federal statutes: 1. the intersecting compliance requirements of the Family Educational Rights and Privacy Act, Children's Online Privacy Protection Act, and Student Online Personal Information Protection Act (FERPA/COPPA/SOPIPA); 2. the Privacy Rule in Health Insurance Portability and Accountability Act (HIPAA). Both sets of statutes are complex in their content and form. Our work in modeling these statutes is motivated by the goal of demonstrating how a computable contract scheme can increase accuracy, efficiency, and consistency in the interpretation of complex legal landscapes, which would be valuable for both lawyers working in those domains and laypeople who are affected by the relevant laws.

4.1 FERPA Prototype

In this prototype, we have modeled with CDL the intersecting compliance requirements of the Family Educational Rights and Privacy Act, Children's Online Privacy Protection Act, and Student Online Personal Information Protection Act (FERPA/COPPA/SOPIPA). The prototype is online available at `http://compk.stanford.edu/ferpa.html`.

The prototype allows interactive formation of an agreement between an information service provider and a district as well as analysis of multiple agreements. It demonstrates the capabilities and added value of our proposed language and reasoning techniques in the situation where an information service provider wishes to obtain data about school going children. In such a case, the information service provider is required by law to enter into a contract with the districts controlling the schools. In the prototype, a contacting party can fill in the details such as which student's data is to be shared, the potential use of the data by the provider, age/grades level of the students etc. The prototype then checks the validity of the contract as per FERPA, COPPA and SOPIPA and displays the violations and obligations if any. The behavior of the user interface is directly determined by FERPA, COPPA and SOPIPA rules modeled with CDL. For the purpose of the demo, we allow users to edit the rules and verify the change in the behavior of the system.

Below we present an excerpt of the database and the rules that we have modeled for this case study. The complete set of rules is visible in the 'Law' tab of the prototype.

4.1.1 Views

The following view definitions define the categories `district_data_pii`. Categories `district_data_non_pii`, `additional_data_pii` and `additional_data_non_pii` can be defined analogously.

```
district_data_pii(D,district_student_name) :-
   district_data(D,district_student_name)

district_data_pii(D,district_student_
   parent_name) :-
   district_data(D,district_student_parent_name)

district_data_pii(D,district_student_dob) :-
   district_data(D,district_student_dob)

district_data_pii(D,district_student_address):-
   district_data(D,district_student_address)

district_data_pii(D,district_student_ssn):-
   district_data(D,district_student_ssn)

district_data_pii(D,district_student_moms_
   maiden_name):-
   district_data(D,district_student_moms_
   maiden_name)

district_data_pii(D,district_student_pob):-
   district_data(D,district_student_pob)
```

The following view definition states a provider is under direct control of district if the provider can amend terms with consent. Other views can be modeled analogously.

```
provider_under_direct_control_of_district(D) :-
   provider_can_amend_terms_with_consent(D)
```

For the complete set of view definition we refer to the ‘Law’ tab of the prototype.

4.1.2 *Constraints*

Below the CDL modeling of the constraint that for any district data PII, if the selected FERPA provision is school official exemption, then the provider must be under direct control of the district. Other constraints can be modeled analogously.

```
illegal("Provider must be under direct control
   of District.") :-
   district_data_pii(D,A) &
   ferpa_provision(D,school_official_exemption) &
   ~provider_under_direct_control_of_district(D)
```

The following rule states the provider must select a FERPA provision.

```
illegal("Must select FERPA exemption.") :-
   district_data_pii(D,A) &
   ~ferpa_provision(D,actual_parental_consent) &
   ~ferpa_provision(D,directory_exemption) &
   ~ferpa_provision(D,school_official_exemption)
```

The following rule states that commercial use of data is prohibited under school official exemption.

```
illegal("Under School Official Exemption,
   commercial use of data is prohibited."):-
   ferpa_provision(D,school_official_exemption)
   district_potential_use_by_provider(
   D,district_4aiii)
```

The following rule states that if the district is in California, then commercial use of any data from the district is prohibited.

```
illegal("SOPIPA prohibits commercial use of
   any data.") :
   district_in_california(D) &
   district_data_pii(D,A) &
   district_potential_use_by_
   provider(D,district_4aiii)
```

For the complete set of modeled view definition we refer to the ‘Law’ tab of the prototype.

4.1.3 *Obligations*

Below the CDL modeling of the consequence that if FERPA provision is directory exemption, then the district must allow opportunity for parent to opt-out of the disclosure of student’s data.

```
add(D,"District must allow opportunity for
   parents to opt-out of the disclosure of
   student+data.") :-
   district_data_pii(D,A) &
   ferpa_provision(D,directory_exemption)
```

For the complete set of modeled obligations we refer to the ‘Law’ tab of the prototype.

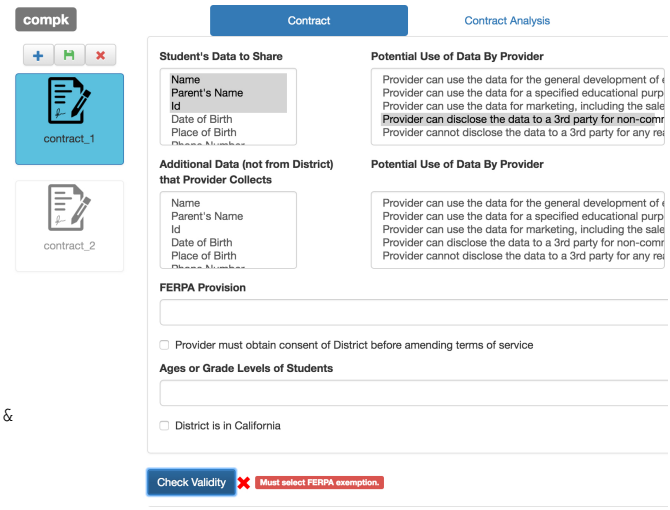


Figure 1. Screenshot of our prototype showing that the contract is invalid as well as the reason for the invalidity.

Validity Checking As briefly mentioned above, our prototype can automatically check whether a contract is valid according to a law. This feature can be very useful in the contract formation phase. Figure 1 shows an example contract in which the information service provider is not under the control of the district, and therefore the selected data artifacts to be shared for the selected intended use is not allowed by the law. In the example contract shown in Figure 1, the &violation is computed because of the presence of the following rule.

```
illegal("Provider must be under direct control
   of District.") :- district_data_pii(D,A) &
   ferpa_provision(D,school_official_exemption) &
   ~provider_under_direct_control_of_district(D)
```

In addition to computing whether a contract is valid or not, our prototype can also automatically output the reason for the invalidity or any rights, limitations and obligations that parties have if the contract is valid. In our example, if the FERPA is changed to ”Actual Parental Consent”, then the contract becomes valid, and our prototype can also automatically compute that the district may disclose the data only according to the terms of the parental consent.

Hypothetical Analysis In addition to validity checking as described above, the FERPA prototype also supports hypothetical reasoning over a set of (valid) contracts. Typically, an information service provider enters into multiple contracts, one for each district, to be able to achieve broader coverage for his/her service. Given a set of such contracts, an information provider is often faced with the problem of deciding whether he/she may use certain data artifact for a particular use. In order to obtain the answer to such a question with our prototype, the information service provider would formulate his question as a query in the ‘Contract Analysis’ tab. Our prototype then analyses all the existing contracts of the information service provider and produces the answer to the question. For example, if an information service provider wishes to know which data he/she may share with a third party, he/she would pose the query `provider_may_share(District,Data,3rd_party,Use)`. The answer to this query will contain all (district, data artifact, usage) that the provider may share with a 3rd party (see also Figure 2).

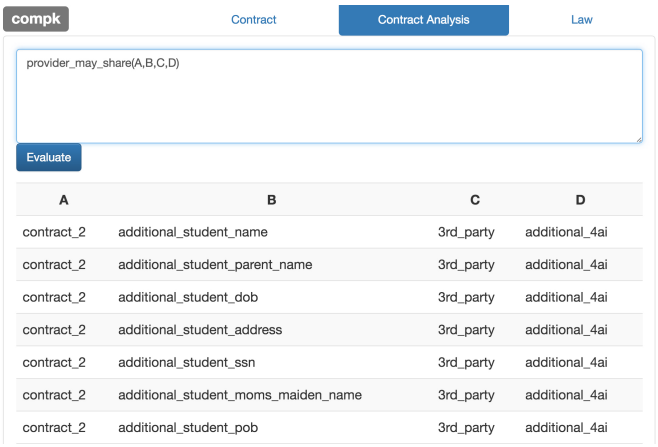


Figure 2. Screenshot of the output of hypothetical analysis

4.2 HIPAA Prototype

In our HIPAA prototype, we specifically modeled the Privacy Rule of the statute. Under the HIPAA Privacy Rule, all types of healthcare providers and related organizations, collectively known as 'covered entities,' are the main subjects of regulation and must comply with the Rule's provisions when deciding if and how to disclose a patient's information, called Protected Health Information (PHI). The Rule's complexity can hinder compliance and lead to potentially huge penalties for parties that violate the Rule. Thus, providing clarity using a computational contract scheme can be immediately valuable to all parties involved. Because the core of HIPAA privacy compliance is what covered entities can and cannot do with a patient?s PHI, we modeled a couple of situations that resemble hypothetical analysis and validity checking, where using a computational contract scheme can help covered entities navigate their legal options. The prototype is available at `http://compk.stanford.edu/hipaa.html`.

Scenario 1 Covered Entity, X, wants to disclose patient Y's information to third party Z, who is a healthcare startup that wants to market its products to Y. Under HIPAA, such disclosure is only legal if Y provides explicit written authorization in plain language for X to issue such disclosure. Any type of non-written authorization would not be valid. In this situation, in order for the covered entity to comply with this constraint, our prototype provides only the option of disclosing PHI for marketing purposes available for X, if a written authorization is issued by Y, allowing X to analyze what are its legal courses of action. This dynamic is modeled by the following rules:

```
legal (X, market(Phi, Z), N) :-
   true(written_authorization(Y, X), M) &
   true(plain_language(Y, X), M) &
   phi (Y, Phi) & ce(X) &
   thirdparty(X, Z) &
   successor (M, N)

true (plain_language(Y, X), M) :-
   does (Y, write_plain_lang(X), M)

true (plain_language(Y, X), N) :-
   true (plain_language(Y, X), M) &
   successor (M, N)

true (written_authorization(Y, X), M) :-
does (Pa, write_authorization(X), M)

true (written_authorization(Y, X), N) :-
   true (written_authorization(Y, X), M) &
   successor(M, N)

legal (X, market(Phi, Y), M) :-
   step(M) &
   does (X, exceptcomm(Y, exception), M) &
   ce(X) & phi(Y, Phi) & successor
```

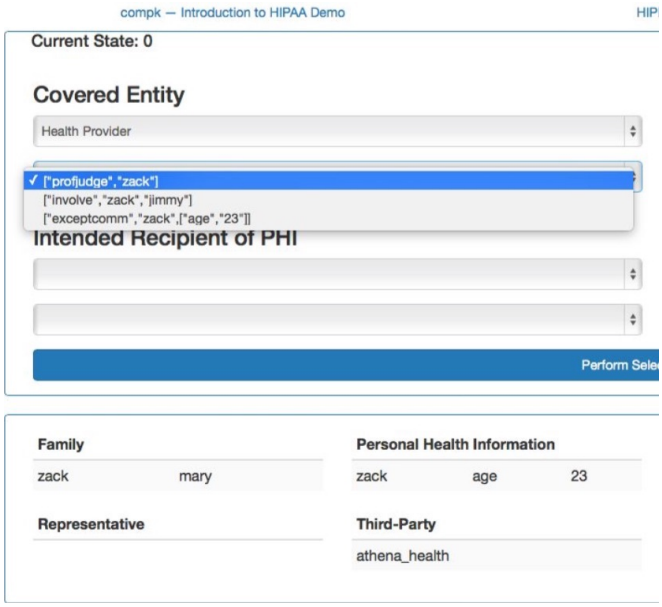


Figure 3. Screenshot showing a health provider's options when it does not have a written permission of a patient

Figure 3 illustrates the scenario, in which the three actors are Zack (the patient), Kantor (Zack's health provider), and Athena (a third party, healthtech startup). Kantor wants to disclose Zack's age, which is a type of PHI specified in HIPAA, to Athena for marketing purposes. In order for Kantor to do this legally, Zack must provide written authorization to Kantor, permitting this disclosure. (Note: non-written permission would not work.) Thus, in state 0, the beginning, when Kantor has not obtained Zack's written permission, Kantor?s options do not include 'marketing'.

However, if Zack does provide written authorization in state 0 (in our product, you would select Zack in the patient box, select the "writtenauth" option under Zack, then click "Perform Selected Actions"), it will move things to state 1, the next state, and Kantor will have the marketing option available to disclose to a third party, like Athena (see Figure 4)

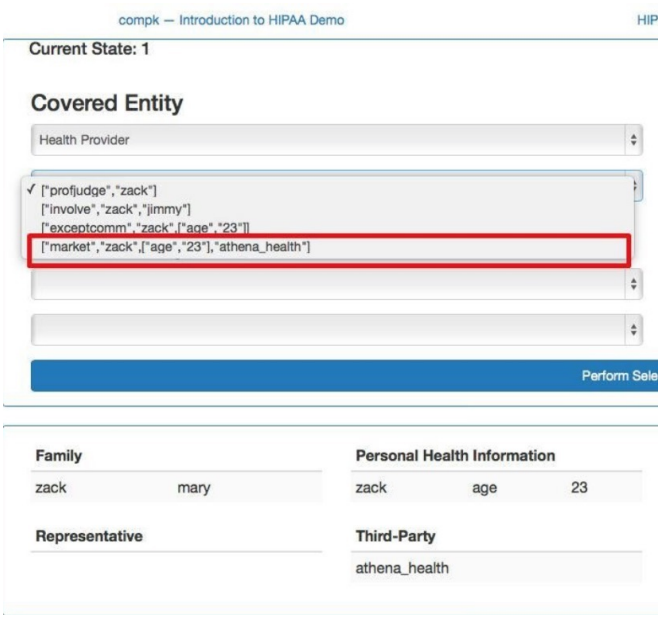


Figure 4. Screenshot showing a health provider's options when it has a written permission of a patient

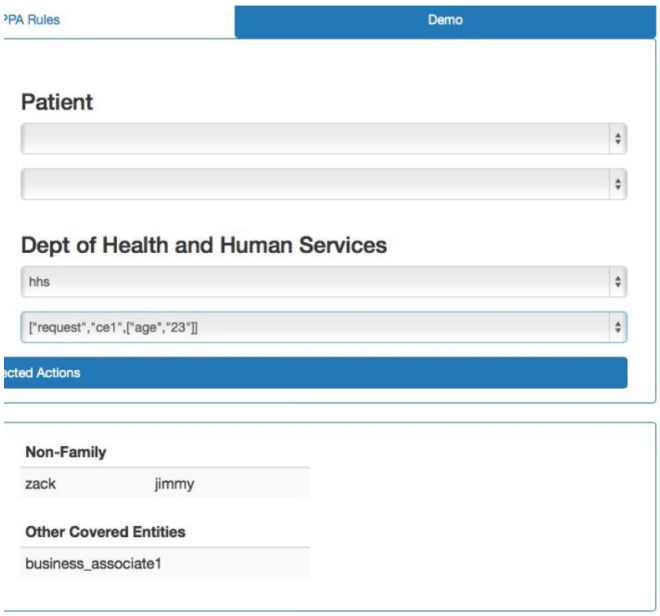


Figure 5. Screenshot showing HHS requesting a health provider to disclose a patient's age.

Scenario 2 this situation models an interaction between Covered Entity, X, and the U.S. Department of Health and Human Services (HHS), which is the main federal agency that regulates and enforces HIPAA compliance. Under the law, if HHS requests PHI from a covered entity, regarding say patient Y, for the purpose of investigating a compliance case, the covered entity must disclose the information to HHS. To depict this constraint in our demo, as soon as HHS issues a request for PHI, X's options are limited to just one: disclose. In fact, if X chooses any other option, our system will conduct validity checking and not allow the step to proceed. Instead, it will immediately alert X that it must disclose the requested PHI. This dynamic is modeled by the following rules:

```
illegal ("Must Disclose PHI") :-
   true(requestment(HHS, X, Phi), M) &
   true(investigation(HHS, X), M) & ce(X) &
   ~does (X, disclose (Phi, HHS), N) &
   phi(Y, Phi) & successor (M, N)

illegal ("Must Disclose PHI")  :-
   does(Y, request(X, Phi), M) & ce(X) &
   phi (Y, Phi) & successor (M, N)&
   ~does (X, disclose (Phi, HHS), N)

illegal ("Must Disclose PHI")  :-
   does (Rep, request(X, Phi), M) &
   phi (Y, Phi) & ce(X) &
   rep (Y, Rep) & successor (M, N) &
   ~does (X, disclose (Phi, HHS), N)
```

This scenario illustrated in Figure 5 is between Kantor and HHS. In HIPAA, when HHS requests PHI from a covered entity like Kantor, for example, if HHS received complaints by patients regarding potential HIPAA violations and wanted to investigate Kantor, Kantor is required to disclose the requested information to HHS, which limits Kantor's legal options. Here in state 2, HHS is requesting Kantor to disclose Zack's age as a part of its investigation.

After HHS commits this request, we move to state 3, where Kantor is legally obligated to disclose Zack's age to HHS. If Kantor does not disclose and instead chooses to commit another action, it is considered illegal; therefore, our system will show an alert stating that PHI must be disclosed and will not move on to the next state until Kantor commits the correct, legally required action (see Figure 6).

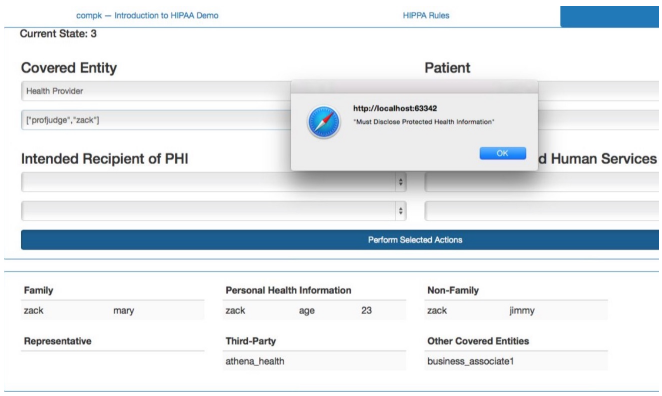


Figure 6. Screenshot showing the enforcement of a health provider's obligation to disclose a PHI of a patient.

5 Related Work

One of the first approaches for formalizing regulations with Logic Programs was presented in [13]. One of the major differences between CDL and the approach presented in [13] is that with CDL one can also express a multi-actor dynamic system and constraints over multiple states.

The formalism presented in [2] can reason over data types (but not over data values) of single individuals (but not of groups), and cannot express certain norms precisely due to lack of parameterized roles. A formalism called pLogic presented in [7] does not allow reasoning over history and future possibilities of actions of various actors. In contrast, CDL can reason over types and (complex logical) interdependencies of objects. CDL can also express and reason over past states and possible actions of an actor in a state in the future. Our formalization technique is also more general than the one presented in [3] as the latter can express only finite state automata. The formalisms presented in [**?**] can describe contracts in terms of deontic concepts such as obligations, permissions and prohibitions. CDL does not directly have special constructs for deontic modalities. However, since deontic modalities can be reduced to temporal modalities and CDL can express dynamic constraints, it is possible to express deontic modalities with CDL.

The term Computable Contracts has been used with various different meanings in the past. In some cases it refers to computer-aided contract formation tools to support a legal professional in drafting a contract in natural language. Such softwares range from simple extensions to popular text processing systems to collaborative web-based contract drafting such as Beagle [6], ClauseMatch [7], and Nitro [8] or more efficient contract readers such LegalSifter [9]. In some other cases such as Kira [10] and LegalRobot [11], natural language contracts are analyzed by extracting terms from contracts. Such analysis techniques are statistics based and cannot do reasoning over complex logical interrelationships of artifacts appearing in a contract. Since such a reasoning capability is a prerequisite for reasoning over dynamics described in a contract, such techniques can also not reason over dynamic constraints on the behavior of the contracting parties.

Recently, the so-called Smart Contracts in Blockchain 2.0, e.g. Ethereum [6, 10], have received a lot of attention [9]. An Ethereum Blockchain is essentially a state-transition system that is managed without central control and secured with cryptographic techniques. Currently, Ethereum Smart Contracts are mostly programmed in a procedural language called Solidity, and thus suffer from the already mentioned problems of procedural hard-coding of contracts. CDL on the other hand is declarative as well as allow for usage of shared vocabulary (e.g. a domain ontology) in multiple contracts. Since the execution semantics of CDL is based on a mapping to state-transition system, we believe that CDL lends itself as a promising alternative to Solidity.

6 Conclusion and Outlook

Our FERPA and HIPAA prototypes have demonstrated two core strengths of making law computable: consistent accuracy and ease of use. Instead of relying on lawyers, whose knowledge on specific subject matters may vary, a service backed up by computable contracts encoded with accurate information can consistently lay out the available legal options for organizations who need help planning their actions, consequently increasing access to justice in particular regulatory areas. This service is also easy to use, for both laypeople and lawyers, because the options being laid out are essentially translated from legalese to plain language that people can understand without prior training. If this type of computable contracts services were to expand, it would not only be valuable to organizations who operate in highly regulated industries, like the covered entities in HIPAA, but also regulatory agencies of these industries, who are in charge of sifting through hundreds of thousands of compliance complaints. For example, from 2003-2013, the number of HIPAA violation complaints sent to HHS increased by more than 900%, and approximately 80% of these complaints are illegitimate, due to simple definitional reasons, e.g. whether the alleged organization falls under the category of covered entities. An extended version of our HIPAA prototype can provide immediate and significant efficiency for HHS, or similar agencies in other countries, to process these types of faulty claims.

One limitation we do acknowledge is our formalization technique's inability to capture inherent nuances that exists in complex legal frameworks. While we assume in our prototypes that every concept has static legal meaning, in reality, many concepts are open to interpretation. This ambiguity often exists by design, because a good piece of law must both incorporate issues of the present day and remain relevant with new behaviors that arise in the future. It is this ambiguity in law that gives rise to disputes and litigations, requiring the input of experienced lawyers and policymakers, who could better anticipate how certain ambiguities would be treated by legislators or judges who have the power to decide what they should mean. This challenge is not something a computable contract scheme is suited to solve, though we can easily envision a future where computable contracts can provide a solid baseline understanding that can be complementary to the work of experienced lawyers, regulators, and lawmakers.

Another area that we will continue to explore is the intersection between computable contracts and the field of Human-Computer Interactions (HCI). With the right kind of user-centric, front-end design, the strength of the expert systems encoded in a computable contract scheme can be easily accessed by ordinary people who need the information contained in these systems, thus significantly enhancing the usability of computable contracts. A complementary partnership between a sleek, intuitive front-end design powered by HCI-principles, and a robust, adaptive back-end system powered by computable contracts could unlock unimaginable benefits for improving access to justice in all societies.

In another thread, we plan to implement a compiler to translate our declaratively specified computable contracts to the EtherScript, the assembly language of Ethereum Virtual Machine while preserving the execution semantics of the contracts. This would enable domain experts and other legal professionals to draft and execute their Smart Contracts themselves instead of getting them programmed by a software developer who may not be a domain expert.

[6] http://beagle.ai
[7] http://clausematch.com
[8] https://www.gonitro.com
[9] https://www.legalsifter.com
[10] https://kirasystems.com/
[11] https://www.legalrobot.com

REFERENCES

[1] K. Adams. Bringing innovation to the tradition of contract drafting: An interview with ken adams. http://blog.scholasticahq.com/post/98149002813/bringing-innovation-to-the-tradition-of-contract, 2014.

[2] Datta A. Mitchell J.C. Barth, A. and H. Nissenbaum. Privacy and contextual integrity: Framework and applications, 2006.

[3] M. D. Flood and O.R. Goodenough. Contract as automaton: The computational representation of financial agreements, 2015.

[4] E. Gerding, 'Contract as pattern language', *Washington Law Review*, **88**(1223), (2013).

[5] Malik Ghallab, Dana S. Nau, and Paolo Traverso, *Automated planning - theory and practice*, Elsevier, 2004.

[6] Evans Jon. Vapor no more: Ethereum has launched. http://techcrunch.com/2015/08/01/vapor-no-more-ethereum-has-launched/ (retrieved May 2016),.

[7] Mitchell J.C. Lam, P.E. and S Sundaram, 'A formalization of hipaa for a medical messaging system', in *In Proc. of the 6th International Conference on Trust, Privacy and Security in Digital Business (TRUSTBUS 2009)*, (2009).

[8] A.V. Lisachenko, 'Law as a programming language', *Review of Central and East European Law*, **37**, 115–124, (2012).

[9] Swan M., *Blockchain: Blueprint for a New Economy 1st Edition*, O?Reilly Media, 2015.

[10] Arvind Narayanan, Bonneau Joseph, Felten Edward, Miller Andrew, and Goldfeder Steven. Bitcoin and cryptocurrency technologies. https://d28rh4a8wq0iu5.cloudfront.net/bitcointech/readings/princeton_bitcoin_book.pdf?a=1 (Retrieved May 2016).

[11] Haapio H. Passera, S. and T. Barton. Innovating contract practices: Merging contract design with information design, 2013.

[12] S. Peppet, 'Freedom of contract in augmented reality: The case of consumer contracts', *UCLA Law Review*, **59**(676), (2012).

[13] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory, 'The british nationality act as a logic program', *Commun. ACM*, **29**(5), 370–386, (May 1986).

[14] H. Smith, 'Modularity in contracts: Boilerplate and information flow', *Michigan Law Review*, **104**(5), 1175–1222, (2006).

[15] H. Surden, 'Computable contracts', *UC Davis Law Review*, **46**(629), (2012).

[16] G.G Triantis, 'Improving contract quality: Modularity, technology, and innovation in contract design', *Stanford Journal of Law, Business, and Finance*, **18**(2), (2013).

Value-Based Reasoning and Norms

Trevor Bench-Capon [1]

Abstract. In this paper we explore how value-based practical reasoning relates to norms and their evolution. Starting from a basic model of a society and the norms that can arise from it, we consider how additional values, extended state descriptions and finer grained descriptions of actions, can lead to more complex norms, and a correspondingly more complex social order.

1 Introduction

Norms are a topic of considerable interest in agents systems [60], [51], [62], [58], [50], [49], [3], [54], [39]. In particular, in open agent systems, it is not possible to assume that all agents will behave according to the same ethical code, and the open nature of the system means that the designer cannot simply impose norms that can be assumed to be followed by all. Of course, it is possible to construct so-called *regulated* systems, where the agent can only perform permissible actions (e.g.[28], [58], [2]). However, since, unlike norms found in legal and moral systems, such norms cannot be violated, it can be argued that (e.g. [35], [30]) they should not be seen as norms all, because the agents have no choice beyond compliance or non-participation. Such rules are thus like the rules of a game, not moral and legal norms.

An excellent starting point for considering the emergence of norms is [57], which does, of course, considerably pre-date multi agent systems, but none the less contains many relevant considerations. In that work, Ullmann-Margalit uses simple two player games, such as the prisoner's dilemma (PD) [46], to explore the topic. In such games there are two players and each can cooperate or defect, and the choices determine the payoffs. In PD as used in [57] mutual cooperation gives a payoff of 3 to each player and mutual defection 1 to each player, while if the actions differ the defector receives 5 and the cooperator receives zero. Some key results concerning PD are that the Nash Equilibrium [48] is where both defect (since defection is the *dominant* action, and will receive the better payoff whatever the other player does) and that a successful strategy in iterated PD (where the players play one another repeatedly) is *Tit-Fot-Tat* [11] (but see [18]). Using *Tit-Fot-Tat* an agent will cooperate in the first round, and then copy its opponent's previous move in every subsequent round. Importantly PD is non-zero sum game: the aggregate utility of mutual cooperation is greater than any other payoff, and the equilibrium in fact yields the lowest collective utility. Thus, if would in fact be mutually beneficial if one offered a payment to the other if they cooperated: this could secure payoff of 3 and 2, so that both would gain over mutual defection. Such agreements are, however, not possible in the game, which does not allow for prior negotiations.

Public goods game have formed the basis of several studies of the emergence of norms in multi-agent systems such as [51], [50], [53], [17], [54] and [39]. An alternative approach is to model a situation as a State Transition Diagram (STD), and to investigate how norms can emerge from agent interaction in such situations [62], [3]. In these latter models, agents are typically represented using the Belief-Desire-Intention (BDI) model [45], [61], inspired by [20]. The BDI model supposes agents to have a set of *beliefs* and a set of dispositional goals (*desires*). Actions are chosen by identifying the desires than can be realised in the current situation (candidate *intentions*), and then committing to one or more of these intentions, and choosing a course of action intended to realise the associated goals. This, however, leaves open the question of where the desires come from in the first place.

Empirical studies suggest, however, that public goods games do not provide a very realistic model of actual human behaviour. Experiments using the public goods games are very common and have formed the subject of metastudies. For example [27] examined 131 examples of the Dictator Game and [42] was based on 37 papers reporting Ultimatum Game experiments. In none of these many studies was the canonical model followed. Although the metastudy of [33] was smaller, looking at only 15 studies, it is particularly interesting in that the studies considered highly homogeneous societies. In BDI systems, there is no explanation of where goals come from. Often they are completely fixed, and even systems where they can be derived from the current state [44], there is a fixed set of potential desires some of which are active in a given situation.

An alternative approach to action selection (often called practical reasoning [47]) is provided by Value-Based Reasoning, in which agents are associated with a set of social *values*, the aspirations or the purposes an agent might pursue, such as liberty, equality, fraternity, wealth, health and happiness, and these values provide reasons why certain situations are considered goals by the agent. The basic idea is that agents have a set of such values and their aspirations and preferences are characterised by their ordering of these social values. Acceptance of an argument as to what to do depends not only on the argument itself - for it must, of course, be a sound argument - but also on the audience to which it is addressed [43]. This notion of audience as an ordering on values was computationally modelled in [31] and made more formal in Value-Based Argumentation Frameworks (VAFs) [12]. VAFs are an extension of the abstract Argumentation Frameworks (AFs) introduced in [24], but whereas in an AF an argument is defeated by any attacking argument, in a VAF an argument is *defeated for an audience* by an attacker only if the value associated with the attacking argument is ranked at least as highly as the attacked argument by that audience. In this way different audiences will accept different sets of arguments (preferred semantics [24] is used to determine acceptance), and, as is shown in [12], provided the VAF contains no cycles in the same value, there will be a unique non-empty preferred extension.

Use of VAFs provides a way of explaining (and computing) the

[1] Department of Computer Science, University of Liverpool, UK. email: tbc@csc.liv.ac.uk

different arguments accepted by different audiences. Value-Based Reasoning been used as the basis of practical reasoning in, amongst others, [29], [6], and [59], and applied in particular areas including law [13], e-democracy [22], policy analysis [55], medicine, [9], experimental economics [14], and rule compliance [21]. Complexity results for VAFs were established in [25] and [41]. Here we will discuss norms and their evolution in terms of the Value-Based approach to practical reasoning.

2 Background

In this section we provide some essential background: the structure with which we use to model our "world", Alternate Action Based Transition Systems (AATS), and the valued-based arguments that agents can use to justify their actions in this environment; the running example we will use to instantiate our model; and the three types of ethical theory we will consider.

2.1 Alternate Action Based Transition Systems

Based on Alternating Time Temporal Logic [4], AATS were originally presented in [62] as semantical structures for modelling game-like, dynamic, multi-agent systems in which the agents can perform actions in order to modify and attempt to control the system in some way. As such they provide an excellent basis for modelling situations in which a set of agents are required to make decisions.

The definition in [62] is:

Definition 1: AATS. An *Action-based Alternating Transition System* (AATS) is an $(n + 7)$-tuple $S = \langle Q, q_0, Ag, Ac_1, \ldots, Ac_n, \rho, \tau, \Phi, \pi \rangle$, where:

- Q is a finite, non-empty set of *states*;
- $q_0 \in Q$ is the *initial state*;
- $Ag = \{1,...,n\}$ is a finite, non-empty set of *agents*;
- Ac_i is a finite, non-empty set of actions, for each $ag_i \in Ag$ where $Ac_i \cap Ac_j = \emptyset$ for all $ag_i \neq ag_j \in Ag$;
- $\rho : Ac_{ag} \rightarrow 2^Q$ is an *action pre-condition function*, which for each action $\alpha \in Ac_{ag}$ defines the set of states $\rho(\alpha)$ from which α may be executed;
- $\tau : Q \times J_{Ag} \rightarrow Q$ is a partial *system transition function*, which defines the state $\tau(q, j)$ that would result by the performance of j from state q. This function is partial as not all joint actions are possible in all states;
- Φ is a finite, non-empty set of *atomic propositions*; and
- $\pi : Q \rightarrow 2^\Phi$ is an interpretation function, which gives the set of primitive propositions satisfied in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q.

AATSs are particularly concerned with the joint actions of the set of agents Ag, J_{Ag}: j_{Ag} is the joint action of the set of n agents that make up Ag, and is a tuple $\langle \alpha_1,...,\alpha_n \rangle$, where for each α_j (where $j \leq n$) there is some $ag_i \in Ag$ such that $\alpha_j \in Ac_i$. Moreover, there are no two different actions α_j and $\alpha_{j'}$ in j_{Ag} that belong to the same Ac_i. The set of all joint actions for the set of agents Ag is denoted by J_{Ag}, so $J_{Ag} = \prod_{i \in Ag} Ac_i$. Given an element j of J_{Ag} and an agent $ag_i \in Ag$, ag_i's action in j is denoted by j^{ag_i}. This definition was extended in [6] to allow the transitions to be labelled with the values they promote.

Definition 2: AATS+V. Given an AATS, an AATS+V is defined by adding two additional elements as follows:

- V is a finite, non-empty set of values.
- $\delta : Q \times Q \times V \rightarrow \{+, -, =\}$ is a *valuation function* which defines the status (promoted (+), demoted (–) or neutral (=)) of a value $v_u \in V$ ascribed to the transition between two states: $\delta(q_x, q_y, v_u)$ labels the transition between q_x and q_y with one of $\{+, -, =\}$ with respect to the value $v_u \in V$.

An *Action-based Alternating Transition System with Values* (AATS+V) is thus defined as a $(n + 9)$ tuple $S = \langle Q, q_0, Ag, Ac_1, ..., Ac_n, \rho, \tau, \Phi, \pi, V, \delta \rangle$. The value may be ascribed on the basis of the source and target states, or in virtue of an action in the joint action, where performing that action itself promotes or demotes a value.

2.2 Reasons for Action

The values give agents reasons to perform or not to perform the various actions, based on the argumentation scheme proposed in [6]. A number of such reasons are given in [8] (the "N" suffix denotes reasons not to perform the action: ϕ is a *goal*, which holds or fails to hold in a given state, and which agents may attempt to realise, maintain, avoid or remove).

R1 We should participate in j in q in which ϕ holds to maintain ϕ and so promote v.
R2N We should not participate in j in q in which ϕ holds since it would remove ϕ and so demote v.
R3 We should participate in j in q in which $\neg\phi$ holds to achieve ϕ and so promote v.
R4N We should not participate in j in q in which $\neg\phi$ holds since it would avoid ϕ and so fail to promote v.
R5 We should participate in j in q to ensure ϕ and so promote v. Note that ϕ may be contingently realised or unrealised in q and that, in some variants, the promotion of v might not be immediate, or permanent. This also applies to R5N and R6.
R5N We should not participate in j in q which would ensure $\neg\phi$ and so demote v.
R6 We should participate in j in q to prevent $\neg\phi$ and so promote v. Note that $\neg\phi$ may be contingently realised or unrealised in q.
R6N We should not participate in j in q which would prevent ϕ and so fail to promote v. We suggest that to make the reason worth consideration we should only use variants which prevent ϕ immediately and permanently.
R7 We should participate in j in q in which $\neg\phi$ to enable ϕ to be achieved and v to be promoted on the next move.
R8N We should not participate in j in q in which ϕ which will risk ϕ being removed on the next move which would demote v.
R9 We should participate in j in q because performing j^{ag} promotes v.
R9N We should not participate in j in q because performing j^{ag} demotes v.

Objections to these arguments can be formed by questioning whether the state is as claimed, the consequences of the action will be as specified, whether the goal is realised and whether the value is indeed promoted. The arguments and attacks are then organised in a Value-Based Argumentation framework (VAF) [12] and evaluated according to an ordering on the values. These value orderings will depend on the subjective preferences of the particular audience, and so different agents may choose different actions.

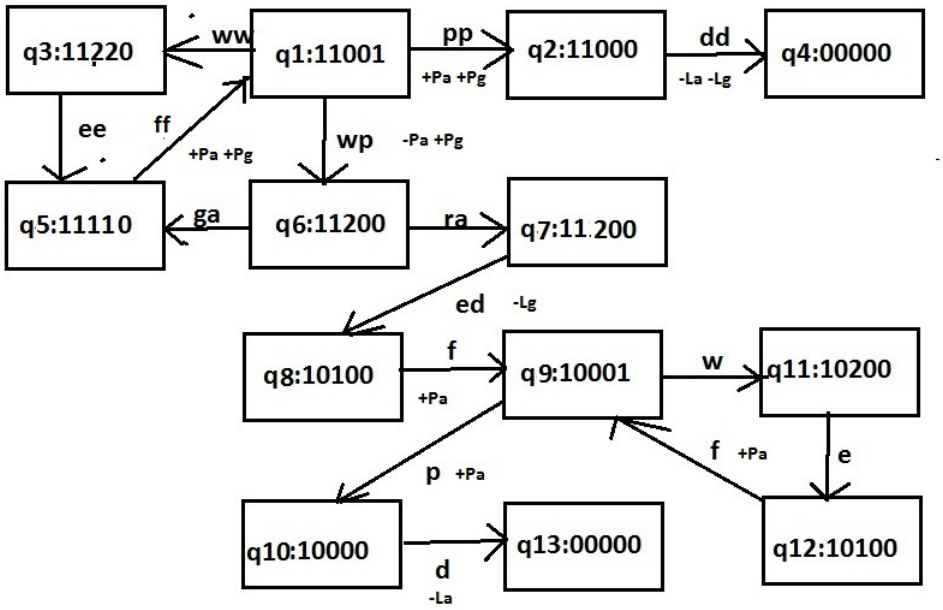


Figure 1. AATS+V for the Example: w = work, p = play, a = ask, g = give, r =refuse, e = eat, f = feast d =die. The same AATS+V is used for both the fable and the parable. Joint actions are ant/father, grasshopper/son. States are: ant/father alive, grasshopper/son alive, ant/father has food. grasshopper/son has food, summer/winter

2.3 Example

An AATS+V was used in [16] to model the states and actions found in both the fable of *the Ant and the Grasshopper* [1] and the parable of *the Prodigal Son* (Luke 15:11-32). Fables and parables are suitable examples for us because they are stories with a moral point. In *the Ant and the Grasshopper*, the story is that during the summer the grasshopper sings and plays while the ant works hard storing up food for the winter. When the winter comes, the grasshopper has no food: nor will the ant give away any of its store, and so the grasshopper dies. In *the Prodigal Son* the prodigal wastes his inheritance on idle play but when destitute asks his father for forgiveness: the father does forgive and takes him back into his household.

An AATS based on the model of [16] is shown in Figure 1. In our example, food is sufficiently abundant in Summer that one can gather food and eat without effort. Growing food for the winter is, however, a full time effort (digging, planting, weeding, reaping, storing) and produces a surplus, but the nature of the activity is that it is either done or not: the amount produced in not proportional to the effort. The food does not last into the summer: therefore the winter ends with a period of carnival (q5, q8 and q12) when the surplus is consumed with feasting. The state has five propositions. The first two indicate whether the ant (father) and the grasshopper (son) are alive. The third and the fourth whether the ant (father) and the grasshopper (son) have no, enough or abundant food, and the fifth whether it is summer or winter. The key decisions are in the initial state (q1) where both grasshopper and prodigal choose to play rather than work and in q6 where the ant refuses the grasshopper (action *r*) while the father gives to the prodigal (action *g*). In the other states there are no choices to be made.

We have labelled the diagram in Figure 1 with just four values. Life for the ant (father) and grasshopper (son) (L_a and L_g) and Pleasure for the ant (father) and the grasshopper (son) (P_a and P_g).

2.4 Ethical Theories

Broadly, as a considerable simplification, ethical theories can be divided into three types:

Consequentialism: An action is right if it promotes the best consequences. For example, Mill's *Utilitarianism* [40].
Deontology An action is right if it is in accordance with a moral rule or principle. For example, Kant [36]
Virtue Ethics: An action is right if it is what a virtuous agent would do in the circumstances. For example, Aristotle [5]

3 Developing a Moral Code

In this section we consider how a moral code might develop from a consideration of value-based practical reasoning in the example scenario.

3.1 Arguments in q_1

We now consider the arguments available to an agent in q_1, based on the values of pleasure and life. The agent's own pleasure and life will be denoted P_s and L_s, the pleasure and life of the other as P_o and L_o. Our arguments are derived from the reasons of section 2.2, expressed in terms of only the agent's own action and the value, e.g. *you should perform* α *since it will promote* v, where α is the agent's action in the justified joint action, and v is the value promoted.

A You should not play since it will risk L_s being demoted (R4N)
B You should work since it will enable P_s to be promoted (R7)
C You should play to promote P_s (R9)
D You should not work since it will demote P_s(R9N)

Thus we have reasons pro and con working: the pro reason is the future pleasure it enables, and the con reason is the immediate loss

of pleasure which accompanies it. Play in contrast affords immediate pleasure, but risks the loss of life. The risk associated with argument A is substantial: avoiding death requires both that the other works, and that the other will forgo its own pleasure in order to save one's life. Therefore (assuming life is preferred to pleasure) only the most risk taking individuals will choose to play in q_1.

Viewed from the perspective of the three moral theories:

- Consequentialism will make work obligatory (or forbid play), to avoid both the undesired consequence of being in q_2 with its unavoidable successor q_4, and the normative collapse that will result from the encouragement given to free loading if the idler is fed [39].
- Deontology will also make work obligatory (or forbid play), since it is not possible to will that both agents play.
- Virtue Ethics will require that life is preferred to pleasure, so that the virtuous agent will choose to work..

All three of these theories will support the continuing existence of the community. We believe a moral code should be *sustainable*, in the sense of ensuring the continuance of the community and avoiding the collapse of its norms [39].

3.2 A first set of moral norms

So let us suppose that the society has the moral norm:

MN1: It is forbidden to play

If all agents act morally the situation can continue indefinitely round the loop q_1, q_3, q_5, q_1. But there will always be the temptation to violate MN1: the value L_s is only threatened, and if q_6 is reached, there is the possibility that that the other agent will give the required food. Work on norms such as that reported in [10] and [39] suggests that unless there is some reinforcement mechanism, norms are liable to break down. The reinforcement mechanism is for violations to be punished by other members of the group, which in q_6 would mean that food is withheld. Moreover, to avoid the norm collapse [39], it is necessary to punish those who do not punish, and so punishment needs to be seen as obligatory. This in turn means that we need a moral norm applicable in q_6:

MN2: It is forbidden to give food

Now refusal to give food would also be justified by an argument based on R6N, *you should not give since that will fail to promote* P_s. Given the counterargument based on R5N, you should not refuse since this will demote L_o, this requires a preference for P_s over L_o. But this does seems selfish rather than moral, and acts against sustainability. It does not seem morally good to prefer a minor value in respect of oneself to an important value in respect of the other: in [7], for example, acting morally was characterised as not regarding lesser values enjoyed by oneself as preferable to more important values in respect of others, which would oblige the agent to give. We can, however, introduce another value, Justice (J), which will justify refusal. This has some intuitive appeal, since the foodless agent has chosen to be in that position, and is attempting to take advantage of the efforts of the other. Thus recognising justice as a third value (labelling the transition q_6-q_7), preferred to L_o, will justify the punishment of violations of MN1. This would be difficult to justify under a consequentialist perspective (since it means the grasshopper dies), but is capable of universal adoption, and it is not difficult to see a preference for justice as virtuous, since it can be justified in terms of equity and sustainability, by preventing the collapse of MN1.

The result will be a pair of norms which are capable of resisting collapse, according to the empirical findings of [39]. The result is a rather puritan society (relieved only by a brief period of hedonism), based on subsistence farming, with a strong work ethic, and an aversion to what in the UK is currently termed a "something for nothing" society. An alternative would be to introduce a fourth value, Mercy, preferred to Justice, and labelling the transition $q6$-q_5. Ranking Mercy above Justice is very possibly the recommendation of the parable of *The Prodigal Son*, and would also allow society to continue, at the expense of a sacrifice of pleasure by the ant. But it is a feature of the parable that the son repents, and there is a tacit understanding that the son will not repeat the pattern, but will choose work in future. We might therefore wish to modify MN2 to something like

- **MN2a** It is allowed to give food only once. We might wish to go further and to accompany this with
- **MN2b** It is obligatory to meet the first request for food

This would represent a preference for Mercy, but enforce *a two strikes and you are out* policy, so that justice is still respected. It also opens the possibility for the ant to play at the grasshopper's expense on some future cycle (cf. children supporting elderly parents). Whereas simply removing MN2 would lead to the possibility of exploitation, and so devalue Justice, the pair of norms MN2a and MN2b retain some respect for Justice, while allowing scope for Mercy until the object of the mercy proves incorrigible. This will require the state vector to have an extra term to record violations.

Our developments beyond the basic scenario of Figure 1 will necessarily be less detailed, both because of space limitations and because of the large number of possible variations. Of course it would, in future, be desirable to extend the scenario at the same level of detail and provide an expanded AATS+V, but we hope that it is clear that the discussion given below in the following sections is making use of the same techniques.

3.3 Critique

Although the norms NM1 and NM2 will give rise to an equitable and sustainable society, we might expect to see thinkers in such a society as questioning the worth of the society. There might be a number of grounds for critiques. For example:

- There is no net pleasure in the society: the displeasure of working is offset by the pleasures of feasting at carnival, but there is no net gain. Such a society lacks progress and reward and any point beyond it own continuance.
- There is no choice or diversity in the society: the path taken is determined at all times by the moral code.
- The pleasure enjoyed in this society is of a rather basic kind, whereas the pleasure it denies itself might be seen as an higher pleasure. The hard line utilitarian might adopt Bentham's view [15] that "Prejudice apart, the game of push-pin is of equal value with the arts and sciences of music and poetry", but others, like Mill would disagree: "it is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied" [40]. Such higher pleasures can only be provided by (certain forms of) play, not by feasting.

Therefore critics might seek a way of changing the moral code so as to improve the society in one or more of these respects. Certainly, it is considered essential to a civilised society that it is able to generate a surplus of food and so allow scope for the development

of arts and sciences, or even the simple enjoyment of leisure time. There is therefore some push behind finding a justification for allowing some relaxation of the rigid morality represented by MN1 and MN2. In order to further the discussion we will distinguish between three types of pleasure, by introducing different values for different pleasurable activities. We will retain P for the bodily pleasures associated with carnival (feasting and the like): toil will also continue to demote this value. We will also distinguish between approved activities made possible by not working (e.g. arts and sciences) which we will term *culture* (C), and deprecated activities (e.g. gaming or mere idleness) which we will term *frivolity* (F). We thus need to distinguish between approved play ($play_a$) (i.e engagement in culture producing activities) and deprecated play ($play_d$) (i.e. engaging in frivolity). We might even modify MN2b to give food only to someone in need because of $play_a$, and to withhold food from those in need as a result of $play_d$.

4 Allowing For Play

There are a number of ways in which we can accommodate players. Some require disparity between agents, while other require a redescription of the world, additional values and a finer grained description of activities and states.

4.1 Power

We first consider a disparity of power. In this situation some agents are sufficiently more powerful than the others to be able to compel them to surrender their food. We assume that the powerful agents comprise less than half the population. This is modelled by allowing the powerful agents to *demand*, rather than *request*, food in q_6, and to render it impossible to refuse a demand, so that there is no *rd* transition between q_6 and q_7. This removes argument A for the powerful, since there is no longer any risk in playing because the demands must be acceded to. Might the powerful play and demand all the food from the ant so that they can also feast? This would result in the ant starving and so would be a short term expedient, since the ant would die and the powerful be forced to work in subsequent years. So we should perhaps have a norm to prevent this:

MN3 It is forbidden to demand non-surplus food.

This can be based on a value preference for the Life of the other over Pleasure.

Remember now that we have distinguished between three types of pleasure so that argument C needs to be split into two arguments:

- **C1**: You should $play_a$ to promote culture (C).
- **C2** : You should $play_d$ to promote frivolity (F).

Now the powerful will not choose to work unless they prefer P to both C and F. They also have a choice of leisure activity, depending on whether they prefer culture of frivolity. Of course, this moral preference is built into the names of the values, and the moral norm, *applicable only to the powerful*, will be

MN4 It is forbidden to $play_d$.

This norm allows the choice to work to be morally acceptable. MN4 is, like Bentham, comfortable with a preference for pleasure over culture. Alternatively we can represent Mill's position with

MN4a It is obligatory to $play_a$

(also directed at the powerful). The problem here is that this means that there is one norm for the powerful and one norm for the powerless. To justify this distinction, there needs to be some kind of social order, recognised by all, so that the class difference between those able to demand in q_6 and those not so able is seen as acceptable. This is not at all unusual in actual societies: for example Mrs Alexander's well known hymn *All Things Bright and Beautiful*, often seen as particularly directed towards children, contains the verse (seldom sung these days):

> "The rich man in his castle, The poor man at his gate, God made them high and lowly And ordered their estate."

This is unproblematic for Consequentialist Theories: indeed given Mill's view that not all pleasures are of equal worth, the consequences are an improvement: since only the powerful *can* act to promote culture, it is good that they do so, even if it is at the expense of the powerless, since culture is preferred to pleasure. Nor is it a problem for Virtue Ethics, since it can enjoin a preference ordering: $L \succ C \succ P \succ F$.

One example of such a society is Feudalism. A good model for such a society is where some agents own the land and allow tenant farmers to work the land in exchange for the payment of rent. The nature of such a society is coloured by the ratio of powerful to powerless. If there are relatively few powerful, they can demand low rents and so leave some surplus to the tenants and allowing some degree of feasting to them (so shortening rather than removing the carnival period). This will also mean that there will be some surplus food available after the needs of the powerful have been met, some of which can be demanded to give the powerful pleasure as well as culture.

What is important, for the sustainability of such a society, is that the powerless respect the social order and do not rise up and overthrow the elite. Revolutions must be avoided. The social order can be reinforced by including a value *deference* (D), promoting by working if one has no power to demand, and by giving when food is demanded, and so promoted by the transitions q_1-q_3 and q_6-q_5. This gives the powerless arguments to respect the social order, to "know their place". Deference can reinforce the preference for C over F by being seen as promoted by the transition using $play_a$ and $work$, but not the transition $play_d$ and $work$ (the idle masters do not command respect). This value recognises two different roles: the powerful are required to promote culture (MN4a) and the rest are required to enable them to do so. Acceptance can be reinforced in several ways including patriotism, in which the powerless are encouraged to take pride in the cultural achievements of their masters; or religion, as in the hymn quoted above. As a further reinforcement, prudence suggests that the rents should not be too high.

A further possibility is that some workers may be taken out of food production and used for other purposes of benefit to all, which might be additional cultural activities (e.g. minstrels), building works (e.g. the pyramids), or whatever, and then fed from the tribute. Thus, once the despot's own needs have been considered, the surplus can be apportioned by them between allowing some retention by its producers and some public works ("bread and circuses"). Oddly the fewer in the powerful class, the greater the scope for ameliorating the lot of the powerless, and hence the society is more likely to be stable. In feudal societies it seems that the powerless suffer more when there is a weak king and squabbling barons rather than when there is a powerful king who keeps the barons in check[2]. The proportion that

[2] For example, in the Robin Hood legends the people favour Richard over his weak brother John.

is taken has been investigated in behavioural economics [38][3]. At the limit, where the classes are equally divided, there is no leeway: there the powerful requires all the surplus.

In addition to Feudalism, there are other models: slavery is one, and the kind of brigandry depicted in the film the *Magnificent Seven* is another. But these afford far less opportunity for keeping the powerless content, and so are liable to breakdown. In the film the banditry is stopped by force, and slavery was abolished, whereas Feudalism evolved into a different social order, rather than being abolished or overthrown (at least in the UK: in France things were ordered differently). The key distinction is restraint on the powerful so that revolution is not seen as worthwhile[4]. To reinforce this, we often find notions of "*noblesse oblige*" or philanthropy. We will term the associated value as *generosity* (G), and it is the *quid pro quo* of deference. This might form the basis of the moral norm:

> **MN5** It is obligatory to be generous in your treatment of the less fortunate

and the virtue ethic ordering: $L \succ C \succ G \succ D \succ J \succ P \succ F$. We still need $C \succ G$ because the point of this social order is to permit $play_a$. G is there to encourage stability, not as an end in itself. Note that, part of this accommodation is to play down which persons actually enjoy the various pleasures. Culture is now seen as a public good and $play_a$ a duty. People are expected to promote the values they can, given their social position. We have accordingly omitted the suffices indicating beneficiaries. Note, however, that generosity could lead the powerless to give away food to the needy: it could replace mercy as a motivation for MN2a and MN2b.

4.2 Wealth

In post-feudal societies we find that class and disparity remain, but that this disparity is manifested as wealth rather than physical coercion. In a sense this transition began in the feudal age, when power began to take the form of (enforceable) land ownership rather than force of arms.

When wealth is the source of power, the forcibly coercive demands of the powerful are replaced by the ability to buy the surplus. So here the transition between q_6 and q_5 becomes buy and $sell$ rather than *ask* (or *demand*) and *give*. In this model, selling is not compulsory and so the possibility of reaching q_7 is there. However not selling restricts the hoarder to promoting P and jeopardises L_o, whereas selling not only avoids demoting L_o, but also opens up the possibility of enjoying some $play_a$ or even $play_d$. For example, by selling half the surplus for two cycles, a worker would be able to save so as to accumulate sufficient wealth to spend the third in play of one or the other kinds and then buy food for the winter. This is the underlying idea of holidays, pensions, and more recently of "gap years". The balance between how the surplus is distributed between $work$, $play_a$ and $play_d$ can be left to the individuals and so made to depend on the preferences of individuals, or there may be norms imposing limits. At his point it is useful to distinguish been values that are maximisers, for which more is always better, and values which are satisficers[5] for which enough can be enough and more is of no benefit and possibly of harm: for example, one will become sated with too much feasting.

In its purest form, this model should lead to a fair degree of equality, since eventually the initially wealthy will have spent all their money, and so be forced to work, since there is no other source of income. There are, however, mechanisms which tend to allow the wealthy to maintain their position:

- The wealthy may own the land (or the means of production) and be in a position to take some proportion of the labour of others in the form of rent or profit. The situation is little different from the feudal, except that payment is now in money, not in kind. The flexibility afforded by money is more suitable to an Industrial society where production requires more than land and labour, and where produce is not bread alone, but a whole range of manufactured goods.
- The wealthy may choose to lend money at interest. Since many will regard a "bird in the hand as worth two in the bush", there is likely be takers for such loans, allowing for people with initial wealth to pay for their needs from the interest and maintain their wealth, and perhaps even, given sufficient borrowers or high enough interest rates, to increase it. Note, however, this requires some way of ensuring that the lenders can be confident that the interest will be paid, and the debt repaid. This in turn requires some kind of norm, e.g.

  > **MN6a** It is obligatory to repay debts.

  This would be associated with a new value of *trustworthiness* or *honesty* (H), promoted by observance of debts (and contracts and agreements generally) and demoted by reneging on such agreements. In order to make this more general we might prefer to use the formalation:

  > **MN6** It is obligatory to honour agreements.

- Some people may have access to wealth from outside. For example, in the sixteenth century, the Spanish rulers had a seemingly inexhaustible supply of gold and silver from the Americas.
- Deference or Generosity may mean that some agents are not required to work or pay but are simply given some kind of tribute. For example monks or priests may be supported by tithes or donations, or the infirm by alms. The latter, where the motivating value is generosity, are perhaps covered by MN4, but this could be modified to be more specific, perhaps as a version of MN5, applicable to all. But the rephrasing as MN5a means that we broaden the notion of *unable to support themselves* from incapacity to include those engaged in some other, worthwhile but unremunerative, activity. This allows us to subsume mercy under generosity, while the qualification still acknowledges justice as a value.

  > **MN5a** It is obligatory to give alms to those unable to support themselves.

  The introduction of honesty may give a value ordering.

  $$L \succ H \succ C \succ G \succ D \succ J \succ P \succ F$$

There is some scope for variation: e.g. P may be ranked higher than J without causing real problems to our moral vision. It is vital

[3] The powerful find themselves in the position of the Dictator in the Dictator Game, or Proposer in the Ultimatum Game. Both of these have been much studied in behavioral economics ([27] and [42]). These studies have suggested that it is rare for people to keep as much as they can for themselves, and that Respondents in the Ultimatum game will take nothing if offered less than what they consider to be a fair amount. Explanations for behaviour in the two games in terms of value-based argumentation can be found in [14].

[4] In the words of the blues song *Custard Pie Blues* by Sonny Terry and Brownie McGhee "You have to give me some of it, or I'll take it all away".

[5] The distinction introduced by Simon [52], although he uses it to describe the attitudes of different people with respect to a single value, namely 'utility'. See also [56] and [14]

that honesty be given such a high ranking as there will normally be reasons based on some other value to break an agreement. Indeed it could be argued that H should even be preferred to L_s since it is always possible (and perhaps desirable) to avoid entering agreements which would risk demoting L_s

We might see a conflict between MN5a and MN2 and its relaxations MN2a and MN2b. In fact what we are doing is recognising a difference between those who cannot work, and whose requests should be granted, and those who could work but choose not to do so[6]. The distinction is intended to enforce MN1, but to allow for some excusable violations (e.g. on the grounds of illness).

4.3 Turn Taking

In the previous subsection we considered situations with an initial imbalance of wealth. But it is possible, given a norm such as MN6, to enable the beneficial trade of surplus production for opportunities for $play_a$, through the mechanism of turn-taking. This arrangement, expressed here as one agent plays this year supported by another agent in return for supporting the play of that agent the following year, is in fact very common as an informal arrangement at the personal level. Many couples or groups living together will come to such an arrangement regarding chores, and the idea of "turn taking" is very common amongst children.

Turn taking also emerged in the empirical work of [37] in which a society of agents played a number of iterated prisoner's dilemma games. The agents had different degrees of *tolerance* (readiness to punish) and *responsiveness* (readiness to cooperate). What emerged was a number of stable situations: mutual cooperation and mutual defection, of course, but also some stable turn taking cycles. These turn taking cycles sometimes benefited the two agents equally, but even where one gained more from the arrangement than the other, it could still be beneficial to both, and to their combined score, when compared with mutual defection. Therefore we might well see such an arrangement emerge, even in an initially equal society, given that C is preferred to P and there is a reinforcing norm such as MN6. As has been noted above, such arrangements are likely to be especially common in domestic situations, where trust is likely to be high. This in turn suggests that it might be possible to differentiate H according to whom it is directed. It is not uncommon to regard it as wrong to cheat family and friends (H_f), dubious to cheat other individuals (H_i), but acceptable (where possible) to take advantage of large ("faceless") organisations (H_o). Such discrimination is rarely enjoined by any ethical theory (although it is possible that, in some circumstances, it would be endorsed by some forms of consequentialism), but is a commonly argued for (and practiced) behaviour. Over-claiming on insurance is not uncommon and is seen by some as a "victimless" crime, suggesting that some might give H_o a very low rank, perhaps even below F.

4.4 Service Provision as Work

In several of of the scenarios discussed previously it came about that because of the preference for C over some other values, certain agents may be enabled to $play_a$ because the consequent promotion of C was such that other agents were inclined to support this activity out of their surplus in preference to P. This is likely to be particularly so in the case pf powerful agents who will choose to act as patrons to certain agents to allow and encourage certain kinds of $play_a$. But similar kinds of patronage may be attractive to other individuals as well, who may be prepared to part with a (typically) small part of their surplus. It is possible that this may emerge with just two agents. The ant may find the singing of the grasshopper so entertaining that he is willing to sacrifice his entire surplus for the privilege of listening to her. But, since the singing of a single grasshopper may entertain a whole colony of ants, it is even more attractive if the cost of supporting the grasshopper can be shared across a large number of individuals. Where this is so, a variety of entertainers can be supported, and other services performed. Money greatly assists this arrangement, and places it on a formal, contractual footing, so that it falls under MN6. As such we might expect the emergence of a service and entertainments sector, where some agents were able to adopt the role of providers of C promoting activities willingly supported by groups of other agents.

This is likely to be increasingly the case when productivity rises, so that workers generate larger surpluses. Now we can adjust our notions of the difference between $play_a$ and $play_d$. We can see $play_a$ as being non-work activities for which people are prepared to pay, and $play_d$ as non-work activities for which people are not prepared to pay. This will require consideration of the agent as well as the activity: people will pay to watch Lionel Messi play football, but no one will pay to watch me play football. We therefore combine norms MN1 and MN4a into single norm:

MN1a It is obligatory to $play_a$ or to $work$.

This differs from MN4 because that norm was directed at only a subset of agents, whereas MN1a can be seen as universal. Interestingly a norm like MN1a may be better supported by a system of reward for $play_a$ rather than punishment for $play_d$. Indeed the payment for the services provided for $play_a$ may well be seen in terms of reward for norm compliance. For a discussion of enforcing norms with rewards rather than punishments see [19].

4.5 Emergence of a State

As well as choosing to spend their surplus on providing themselves with culture, through paying others to $play_a$, agents may choose to pay others to do their duties. In [39] it was shown empirically that to avoid norms collapsing it is necessary that they not only be backed by the punishment of violators, but that those who fail to punish must themselves be punished. Since punishment has a cost, however, there are reasons not to punish, and in societies where violations are comparatively rare, the cost of punishment falls unevenly and unpredictably. We saw how punishment for violating MN1 can naturally be expressed as MN2 (which actually is cost free for the punisher), but when we move to more sophisticated norms such as MN6, punishment may not have a simple manifestation as a norm. Recognising the need to punish is an important aspect of social cohesion: as expressed in [39]:

> This move from enforcement by vigilantes (those taking the law into their own hands) to seeing law enforcement as the social duty of responsible citizens is an important milestone in the development of a society that respects its laws.

Once punishment is seen as a social duty it is a small step to organise and pay for a third party to punish violators. Assuming relatively

[6] The distinction between the deserving and undeserving poor was a central concern of the UK 1834 Poor Law Amendment Act, and is enjoying a revival in popular attitudes expressed in the UK today. It contrasts with the underlying philosophy of the UK Supplementary Benefits Act 1976, which saw a certain minimal level of support as the right of every citizen.

few law breakers a small levy will enable a dedicated agent to be paid to enforce the norms. Of course, non-payment of the levy will also be subject to punishment. From this it is a small step to taxation, and the provision of services such as law enforcement by the state. And if law enforcement, why not other duties? Thus MN5a may be better observed by contribution to a central fund responsible for identifying those who should be supported and providing that support.

In this way States may emerge, first as a Hobbesian *Leviathan* [34], but, once established, available to take on the performance of other duties. Further the State may take on the role of intervention to resolve conflicts of interest between its citizens [23], or to educate its citizens [37]. An emergent State may also lead to new values such as *self-reliance*, *freedom* and *community*, and the relative preferences of these new values, and how they are promoted and demoted by different models of the State may provide insight into the form in which the State emerges. In some circumstances the State may take on an even broader role, and become itself the arbiter of what constitutes $play_a$, by itself supporting certain activities. Thus we often find subsidies for opera, but never for football. Of course, allowing the state to determine what counts as culture in this way will be controversial, and so we may find that we need to to distinguish between two types of $play_a$: high culture as approved by the state and subsidised ($play_{sa}$) and popular culture approved by citizens and paid for out of their own retained surplus ($play_{pa}$). This provides another example of how increasing the level of sophistication of the model necessitates the finer grained discrimination of values and actions.

5 Discussion

As observed by Hare [32], for most people, most of the time, following moral norms involves little more that applying a set of learned principles. Hare, however, also says that there will be occasions when we need to think out a moral problem from fist principles, and that the recognised norms are a useful summary of such reasoning.

> What the wiser among us do is to think deeply about the crucial moral questions, especially those that face us in our own lives, but when we have arrived at an answer to a particular problem, to crystallize it into a not too specific or detailed form, so that its salient features may stand out and serve us again in a like situation without so much thought.

When thinking about the emergence of norms, it is is this deep thinking that gives rise to the norms that we need to model. In this paper we have argued that value-based practical reasoning applied to a model of society expressed as an AATS+V provides the machinery to model this kind of reasoning. Much current work on norm emergence is done using either simulations of public goods games or by proving properties of such games as in [51], or by performing model checking on state transition diagrams as in [62]. The first approach has given some insights, but the simplification necessary, and assumptions about the homogeneity of agents, suggest that there are limitations to the approach. These doubts are strengthened by the fact that the behaviour of people observed empirically in experiments using such games does not support the model used [27] and [42]. The second approach also has a view of agents as highly goal directed, and tends to simplify its representation of norms by removing transitions representing forbidden actions. This means that it is highly effective at proving properties of the system, when the norms are complied with and for verifying the design of norms, but less good in explaining where the norms come from in the first place, and why the agents wish to pursue them, If we are looking for emergence rather than imposition by a designer this is a problem. We believe that the use of value-based argumentation provides a finer grainer account of the reasoning involved, and is therefore better placed to account for the norms that emerge from different social set-ups.

In section 3 we described how two norms might emerge in a simple society. One is a primary norm, the other provides a natural way of punishing transgressions of the primary norm (and a way of removing transgressors). We believe that although the model is simple, it is a not implausible representation of a primitive agricultural society. Subsequently we described how making the model more sophisticated would lead to other norms, and more importantly to the need to introduce additional values (some of which may be *metavalues* promoted and demoted by value orderings rather than actions) and to make finer grained discriminations both in values and in actions. Thus $play$ becomes seen as the socially beneficial $play_a$ and the indulgent $play_d$ and a need to discriminate the value of honesty according to the relationship between the agents involved in the transaction may become apparent. Unfortunately the provision of detailed models, and the particular arguments that they support, is beyond the scope of this workshop paper: all that is possible here is to sketch how additions to the model would result in different norms, and so give a flavour of the process.

We believe that such detailed models would indeed provide a fruitful way of analysing and explaining social developments. Our account here for example, coheres well with the account of social development found in Durkheim [26]. Durkheim suggests that in a "primitive" society people act and think alike with a collective or common conscience, which is what allows social order to be maintained. In such a society laws tend to be highly repressive. Both of these are true of the model presented in section 3, where there is a norm (MN1) to be followed by all and transgressions are effectively punished by death through MN2. Durkheim further argues that in an advanced, industrial, capitalist society, the complex division of labor means that people are allocated in society according to merit and rewarded accordingly, and that diversity is embraced rather than opposed. This accords with our discussion of the norms that develop as surplus production increases, and the development of exchanges enabled by MN6, leading to the increasing prevalence and diversity of service work, rather than food production. Within this framework we could, for example, explore the different norms that emerge when the surplus comes from a general rise in productivity from where it comes as the result of an external boost to wealth, as in sixteenth century Spain. Note also that the sophisticated societies require increased cooperation (supported by norms such as MN6 and values such as trust and honesty) and tended to increase the degree of commercial exchanges between agents. It was these two factors that were found to lead to the greatest deviation from the classical model in the Ultimatum Games studied in [33], supporting the view that the more sophisticated the society the less adequate the model provided by simple public goods game simulations. Thus even if these simulations provide a good account of how initial norms *emerge*, investigating their *development* may require a finer grained approach.

As a final remark we may return to the types of ethical theory mention in 2.4. The consequentialist approach to ethics is reflected in both public goods game simulations which picture agents as homogeneous utility maximisers, and the STD based reasoning of [3] which designates states as desirable and undesirable. In contrast the value-based approach, which allows for agents to have different desires and aspirations represented by their different ordering on values, is more in the virtue ethics tradition. Norms encourage a value order such that agents will want to choose the "right" actions.

REFERENCES

[1] Aesop, *Fables, retold by Joseph Jacobs*, volume Vol. XVII, Part 1, The Harvard Classics. New York: P.F. Collier and Son, 1909-14.

[2] T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge, 'Power in normative systems', in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 145–152, (2009).

[3] T. Ågotnes and M. Wooldridge, 'Optimal social laws', in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pp. 667–674, (2010).

[4] R. Alur, T. Henzinger, and O. Kupferman, 'Alternating-time temporal logic', *Journal of the ACM (JACM)*, **49**(5), 672–713, (2002).

[5] Aristotle, *The Nicomachean Ethics of Aristotle, translated by W.D. Ross*, Heinemann, 1962.

[6] K. Atkinson and T. Bench-Capon, 'Practical reasoning as presumptive argumentation using action based alternating transition systems', *Artificial Intelligence*, **171**(10-15), 855–874, (2007).

[7] K. Atkinson and T. Bench-Capon, 'Addressing moral problems through practical reasoning', *Journal of Applied Logic*, **6**(2), 135–151, (2008).

[8] K. Atkinson and T. Bench-Capon, 'Taking the long view: Looking ahead in practical reasoning', in *Computational Models of Argument - Proceedings of COMMA 2014*, pp. 109–120, (2014).

[9] K. Atkinson, T. Bench-Capon, and S. Modgil, 'Argumentation for decision support', in *Database and expert systems applications*, pp. 822–831. Springer, (2006).

[10] R. Axelrod, 'An evolutionary approach to norms', *American political science review*, **80**(04), 1095–1111, (1986).

[11] R. Axelrod, *The evolution of cooperation*, Basic Books, 1987.

[12] T. Bench-Capon, 'Persuasion in practical argument using value-based argumentation frameworks', *J. of Logic and Computation*, **13**(3), 429–448, (2003).

[13] T. Bench-Capon, K. Atkinson, and A. Chorley, 'Persuasion and value in legal argument', *J. of Logic and Computation*, **15**(6), 1075–1097, (2005).

[14] T. Bench-Capon, K. Atkinson, and P. McBurney, 'Using argumentation to model agent decision making in economic experiments', *Autonomous Agents and Multi-Agent Systems*, **25**(1), 183–208, (2012).

[15] J. Bentham, *The rationale of reward*, John and HL Hunt, 1825.

[16] F. Bex, K. Atkinson, and T. Bench-Capon, 'Arguments as a new perspective on character motive in stories', *Literary and Linguistic Computing*, **29**(4), 467–487, (2014).

[17] C Bicchieri, *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press, 2005.

[18] K. Binmore, 'Review of robert axelrod complexity and cooperation', *Journal of Artificial Societies and Social Simulation*, **1**(1), (1998).

[19] A. Boer, 'Punishments, rewards, and the production of evidence', in *Proceedings of JURIX 2014*, pp. 97–102, (2014).

[20] M.E. Bratman, *Intention, Plans, and Practical Reason*, The David Hume Series, Cambridge University Press, 1999.

[21] B. Burgemeestre, J. Hulstijn, and Y-H. Tan, 'Value-based argumentation for justifying compliance', *AI and Law*, **19**(2-3), 149–186, (2011).

[22] D. Cartwright and K. Atkinson, 'Using computational argumentation to support e-participation', *Intelligent Systems*, **24**(5), 42–52, (2009).

[23] A. Chorley, T. Bench-Capon, and P McBurney, 'Automating argumentation for deliberation in cases of conflict of interest', in *Proceedings of COMMA 2006*, pp. 279–290. IOS Press, (2006).

[24] P. M. Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial intelligence*, **77**(2), 321–357, (1995).

[25] P. Dunne, 'Tractability in value-based argumentation', in *Proceedings of COMMA 2010*, pp. 195–206, (2010).

[26] E. Durkheim, *The division of labor in society*, Simon and Schuster, 2014. First published 1893.

[27] C. Engel, 'Dictator games: a meta study', *Experimental Economics*, **14**(4), 583–610, (2011).

[28] M. Esteva, D. De La Cruz, and C. Sierra, 'Islander: an electronic institutions editor', in *Proceedings of AAMAS 02*, pp. 1045–1052, (2002).

[29] A. Garcez, D. Gabbay, and L. Lamb, 'Value-based argumentation frameworks as neural-symbolic learning systems', *J. of Logic and Computation*, **15**(6), 1041–1058, (2005).

[30] G. Governatori, 'Thou shalt is not you will', in *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pp. 63–68. ACM, (2015).

[31] F. Grasso, A. Cawsey, and R. Jones, 'Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition', *International Journal of Human-Computer Studies*, **53**(6), 1077–1115, (2000).

[32] R. M. Hare, *Freedom and reason*, Oxford Paperbacks, 1965.

[33] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath, 'In search of homo economicus small-scale societies', *The American Economic Review*, **91**(2), 73–78, (2001).

[34] Thomas Hobbes, *Leviathan, 1651*, Scolar Press, 1969.

[35] A. Jones and M. Sergot, 'Deontic logic in the representation of law: Towards a methodology', *AI and Law*, **1**(1), 45–64, (1992).

[36] I. Kant, *Groundwork of the Metaphysics of Morals*, CUP, 1998. First Published 1785.

[37] M. Lloyd-Kelly, K. Atkinson, and T. Bench-Capon, 'Fostering cooperative behaviour through social intervention', in *Proceedings of (SIMULTECH), 2014*, pp. 578–585. IEEE, (2014).

[38] G. Loewenstein, 'Experimental economics from the vantage-point of behavioural economics', *The Economic Journal*, **109**(453), 25–34, (1999).

[39] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, T. Bench-Capon, and M. Luck, 'Establishing norms with metanorms in distributed computational systems', *AI and Law*, **23**(4), 367–407, (2015).

[40] J.S. Mill, *Utilitarianism*, Longmans, Green, Reader, and Dyer, 1871.

[41] Samer Nofal, Katie Atkinson, and Paul E Dunne, 'Algorithms for decision problems in argument systems under preferred semantics', *Artificial Intelligence*, **207**, 23–51, (2014).

[42] H. Oosterbeek, R. Sloof, and G. Van De Kuilen, 'Cultural differences in ultimatum game experiments: Evidence from a meta-analysis', *Experimental Economics*, **7**(2), 171–188, (2004).

[43] C. Perelman, *The new rhetoric*, Springer, 1971.

[44] I. Rahwan and L. Amgoud, 'An argumentation based approach for practical reasoning', in *Proceedings of AAMAS 06*, pp. 347–354, (2006).

[45] A. S. Rao and M.P. Georgeff, 'Modeling rational agents within a bdi-architecture', in *KR 91*, pp. 473–484, (1991).

[46] A. Rapoport and A. Chammah, *Prisoner's dilemma: A study in conflict and cooperation*, volume 165, University of Michigan press, 1965.

[47] J. Raz, *Practical Reasoning*, Oxford University Press, Oxford, 1979.

[48] A.E. Roth and J. K. Murnighan, 'Equilibrium behavior and repeated play of the prisoner's dilemma', *Journal of Mathematical psychology*, **17**(2), 189–198, (1978).

[49] B. Savarimuthu, Maryam Purvis, Martin Purvis, and S. Cranefield, 'Social norm emergence in virtual agent societies', in *Declarative Agent Languages and Technologies VI*, 18–28, Springer, (2008).

[50] S. Sen and S. Airiau, 'Emergence of norms through social learning.', in *IJCAI*, volume 1507, p. 1512, (2007).

[51] Y. Shoham and M. Tennenholtz, 'On the emergence of social conventions: modeling, analysis, and simulations', *Artificial Intelligence*, **94**(1), 139–166, (1997).

[52] H. A. Simon, 'Rationality as process and as product of thought', *The American economic review*, **68**(2), 1–16, (1978).

[53] Brian Skyrms, *Evolution of the social contract*, CUP, 2014.

[54] T. Sugawara, 'Emergence and stability of social conventions in conflict situations', in *IJCAI*, pp. 371–378, (2011).

[55] I. Tremblay, J.and Abi-Zeid, 'Value-based argumentation for policy decision analysis project in québec', *Annals of Operations Research*, **236**(1), 233–253, (2016).

[56] A. Tversky and D. Kahneman, 'Judgment under uncertainty: Heuristics and biases', *science*, **185**(4157), 1124–1131, (1974).

[57] E. Ullmann-Margalit, *The emergence of norms*, Clarendon Press Oxford, 1977.

[58] W. van Der Hoek, M. Roberts, and M. Wooldridge, 'Social laws in alternating time: Effectiveness, feasibility, and synthesis', *Synthese*, **156**(1), 1–19, (2007).

[59] T. van der Weide, F. Dignum, J-J Ch Meyer, H. Prakken, and GAW Vreeswijk, 'Multi-criteria argument selection in persuasion dialogues', in *Argumentation in Multi-Agent Systems*, 136–153, Springer, (2011).

[60] A. Walker and M. Wooldridge, 'Understanding the emergence of conventions in multi-agent systems.', in *Proceedings of ICMAS 95*, pp. 384–389, (1995).

[61] M. Wooldridge, *An introduction to multiagent systems*, John Wiley & Sons, 2009.

[62] M. Wooldridge and W. van der Hoek, 'On obligations and normative ability: Towards a logical analysis of the social contract', *Journal of Applied Logic*, **3**, 396–420, (2005).

Rules are Made to be Broken

Trevor Bench-Capon[1] and **Sanjay Modgil**[2]

Abstract. There is an increasing need for norms to be embedded in technology as the widespread deployment of applications such as autonomous driving and warfare and big data analysis for crime fighting and counter-terrorism becomes ever closer. Current approaches to norms in multi-agent systems tend either to simply make prohibited actions unavailable, or to provide a set of rules (principles) which the agent is obliged to follow, either as part of its design or to avoid sanctions and punishments. We argue that both these approaches are inadequate: in order to meet unexpected situations agents must be capable of violating norms, when it is appropriate to do so, either accepting the sanction as a reasonable price to pay, or expecting the sanction to not be applied in the special circumstances. This in turn requires that agents be able to reason about what they should do from first principles, and one way to achieve this is to conduct value based reasoning using an argumentation scheme designed for practical reasoning. Such reasoning requires that agents have an acceptable set of values and an acceptable ordering on them. We discuss what might count as an acceptable ordering on values, and how such an ordering might be determined.

1 Introduction

As noted in the workshop call for papers, there is an increasing need for norms to be embedded in technology as the widespread deployment of applications such as autonomous driving and warfare and big data analysis for crime fighting and counter-terrorism becomes ever closer. Current approaches to norms in multi-agent systems tend either to simply make prohibited actions unavailable (e.g. [33]) or to provide a set of rules (principles) which the agent is obliged to follow, in the manner of Asimov's Three Laws of Robotics [4]. Neither of these methods can be seen as satisfactory ways of providing moral agents (i.e agents able to reason and act in accordance with norms) since not only is it in the nature of norms that they *can* be violated, but circumstances may arise where they *should* be violated. In fact norms are, in real life and also in MAS, typically backed by sanctions [10]. The idea behind sanctions is to change the consequences of actions so as to make compliance more pleasant and/or violation less pleasant[3]. As noted in [10], sanctions can be seen as *compensation* (like library fines) when they can be viewed as a charge for violation, which makes the situation acceptable to the norm issuer, or as *deterrents*, where the sanctions are meant to ensure compliance by relying on the self-interest of the norm subject. When the norm *should* be violated sanctions may be problematic as they disincentivise the agent. This problem can be lessened in cases where the violation can be condoned and the sanction not applied, but this requires an agreement between the agent and the agent imposing the sanction that the violation was justified (often not the case: consider dissidents such as Gandhi and Mandela). Moreover sanctions need to be enforced, otherwise agents may take the risk of escaping punishment, and violate the norm when there is no acceptable reason to do so.

Thus an important reason for thinking in terms of norms is the recognition that on occasion they need to be violated [24]. While the norm is intended to provide a useful heuristic to guide behaviour, allowing for a quick unthinking response, unreflecting adherence to such moral guidelines is not what we we expect from a genuinely moral reasoner. R.M. Hare, a leading moral philosopher of the last century, expressed it thus [22]:

> There is a great difference between people in respect of their readiness to qualify their moral principles in new circumstances. One man may be very hidebound: he may feel that he knows what he ought to do in a certain situation as soon as he has acquainted himself with its most general features ... Another man may be more cautious ... he will never make up his mind what he ought to do, even in a quite familiar situation, until he has scrutinized every detail. (p.41)

Hare regards both these extreme positions as incorrect:

> What the wiser among us do is to think deeply about the crucial moral questions, especially those that face us in our own lives, but when we have arrived at an answer to a particular problem, to crystallize it into a not too specific or detailed form, so that its salient features may stand out and serve us again in a like situation without so much thought. (p.42)

So while principles may serve well enough most of the time, there are situations where we need to think through the situation from scratch. In this paper we will consider how we can give software agents the capacity to perform quasi-moral reasoning[4].

2 Problems With Current Treatments

There are two main approaches to enforcing normative behaviour in MAS: either by removing prohibited actions (e.g. [33]), or by including explicit rules expressing the norms, often accompanied by

[1] Department of Computer Science, University of Liverpool, email: tbc@csc.liv.ac.uk
[2] Department of Informatics, King's College, London
[3] In decision theoretic terms, the ideal for deterrence being for violations to yield an overall negative utility.

[4] We say "quasi-moral" since software agents do not themselves have ethical status, and cannot be considered to share our values. In this paper we will see such agents as proxies for human beings in simulations or transactions, and so their values will be those of the human they are representing. Developing a set of values applicable to software agents would be the topic of another paper. To see that human values are not applicable to software agents consider the fact that their life is of little value, since they can be easily reproduced or replaced, they don't feel pleasure or pain, nor happiness nor sorrow, and have no experience of liberty or fraternity.

sanctions. Neither are entirely satisfactory. We will illustrate our discussion with a model of the fable of **the** *Ant and the Grasshopper* [1], previously used in [14]. The model takes the form of an Alternating Action-Based Transition (AATS) [33], augmented with value labels [6]. The transition system, in which the nodes represent the states the agent may reach and the actions it may use to move between them (in an AATS they are *joint* actions, one action for each relevant agent), is a typical ingredient of Multi Agent Systems (MAS): the value labelling provides the basis for moral reasoning.

In the fable the ant works throughout the summer, while the grasshopper sings and plays and generally indulges herself. When winter comes and the ant has a store of food and the grasshopper does not, the grasshopper asks the ant for help. The ant refuses and says the grasshopper should have foreseen this, and so the grasshopper starves. The same model also can be used to represent the parable of *the Prodigal Son*, except that in the parable the father welcomes the repentant prodigal back, and does give him food.

Using the first approach we would enforce the behaviour recommended by the fable by removing the transition from q_6 to q_5 or the behaviour of the parable by removing the transition from q_6 to q_7. A real life example in which actions are made unavailable is erecting bollards to prevent vehicles from entering a park (to use the famous example of Hart [23]). What can be wrong with this approach? After all, we can *prove* that the undesirable situation will not be reached, either using model checking [17] or analytic methods. Thus we can prove that universal compliance with the norm will achieve the desired results. This may be so, so long as the situation envisaged in the model is in operation. But suppose some state not modelled arises: perhaps someone has a heart attack in the middle of the park and so it is essential for an ambulance to enter the park in order to save that person's life. Now the bollards will prevent the person from being saved, and the object of the norm, i.e. the value that the norm is designed to serve, the safety of park users, will be demoted rather than promoted. While the norm is effective in an ideal world, we do not live in an ideal world, and in a sub-ideal world it is often the case that adhering to the norms applicable to an ideal world will not lead to the most desirable results[5].

Similarly, principles may cease to prescribe the best course of action in unforeseen situations. The whole point of Asimov's three laws as a fictional device is that following them may lead to outcomes that the principles were designed to avoid. While any set of principles may provide good guidance most of the time, it is not difficult to think of gaps, situations and conflicts where following the principles will lead to undesirable results, and so need to be disregarded. The problem is not improved by the existence of sanctions, and indeed may be made worse since the threat of possible punishment makes violation less attractive to the agent.

Thus while either of the approaches may be effective in closed systems (providing they are simple enough for a model covering every eventuality to be constructed), they cannot be sure to cope with the unexpected events and states that will arise in an open-system, where not every possibility can be envisaged or modelled[6]. In such cases we may find that the very reasons which led to the adoption of a norm will require the agent to violate that very same norm.

Irrespective of which option is chosen, the regulation of behaviours at the level of norms does not allow for agents to appropriately violate norms, in cases where compliance with the normatively prescribed behaviours results in demotion of the values that these norms are designed "to serve", or even of other, preferred, values. Hence, we argue that agents should be equipped with the capacity to reason about values, the extent to which normatively prescribed actions serve these values, which values are more important than other values (i.e. value orderings qua 'audiences'), and the ability to derive these orderings from a variety of sources, including experience, the law, and stories prevalent in the culture. These capacities constitute moral reasoning from first principles; the kind of reasoning required to deal with new and unexpected situations in which blind compliance with norms may lead to undesirable outcomes. This paper serves as a call to further develop reasoning of this kind, building on a number of existing developments that we survey.

[5] This is known in economics as the *Theory of the Second Best* [25].

[6] As Wilde put it in *An Ideal Husband*: "To expect the unexpected shows a thoroughly modern intellect".

3 Value-Based Reasoning

A method for value-based reasoning was proposed in [8], formalised using an AATS labelled with values in [6] and further articulated in [5], and which gave nine reasons for action in terms of the promotion and demotion of values. The basic idea is that the transitions which promote values form the basis of arguments for the action which will allow that transition to be followed, and that the transitions which demote values will supply arguments against actions which permit these transitions. Further arguments may come from assumptions about the current state and the state that will be reached by following a particular transition. These arguments and the attack relations between them (determined according to the so-called critical questions listed in [6]) define an argumentation framework [20]. Moreover since the arguments will be associated with values, the framework is a value-based argumentation framework (VAF) [9]. In a VAF the arguments are evaluated from the perspective of an *audience* (cf [31]) characterised as an ordering on values, and attacks which are unsuccessful for an audience are distinguished from those which succeed (*defeats*). The result is a set of arguments acceptable to a particular audience. If there are no cycles in a single value, this set will be non-empty and unique [9].

If we consider the ant's choice in q_6 of Figure 1, he may either refuse or give. Which is chosen will, using the labels of Figure 1, depend on whether the ant prefers his own pleasure to the life of the grasshopper. The application of value based reasoning to moral decisions was considered in [7], which suggested that moral acceptability required that one's own lesser values should not be more highly ranked than more important values relating to others. This would not (morally) allow the preference of the ant's pleasure over the grasshopper's life, and so require the ant to give food to the grasshopper. But the labelling in Figure 1 is not the only one possible. If we think more abstractly we may see the ant's refusal as promoting *Justice*, since the grasshopper knew full well that food would be required in the winter and not working in the summer would mean later exploitation of the good nature of the ant. Similarly we could label the giving of the food as *compassion* or *mercy*. Preferring justice to mercy becomes more legitimate if we consider the role of the moral code to be producing a sustainable society, which requires that working in the Summer be seen as the norm. As shown in [27] the sustainability of norms requires that transgressions be subject to punishment, and so punishing the grasshopper may be seen as the duty of the ant. Note too that in the parable the prodigal is repentant, and so the father will only be expected to show compassion once. Representing such things as repentance will require an extension to the state descriptions to record histories, but will allow a preference for justice over compassion to be dependent on the misbehavior being repeated. Benefits of tolerance of limited misbehaviour before en-

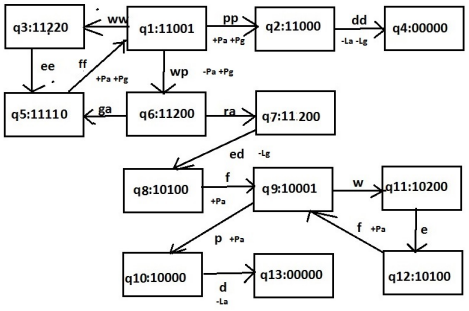


Figure 1. AATS+V: w = work, p = play, a = ask, g = give, r =refuse, e = eat, f = feast d =die. The same AATS+V is used for both the fable and the parable. Joint actions are ant/father, grasshopper/son. States are: ant/father alive, grasshopper/son alive, ant/father has food. grasshopper/son has food, summer/winter

forcing punishments is explored through simulation in [26].

Yet another way of describing the problem would be to recognise that the singing of the grasshopper may be a source of pleasure to the ant as well as to the grasshopper. Seen this way, the ant does not so much give food to the grasshopper as to pay for services rendered. This in turn requires requires recognition that it is the duty of the ant to pay for the services of the grasshopper, and so justice is now promoted by following the transition from q_6 to q_5, not q_7 . Moreover since a single grasshopper may entertain a whole colony of ants, the burden falling on a single ant may be relatively small.

If, however, there is only a single ant, suppose that the harvest fails, and there is no surplus to pay the grasshopper. Should the ant follow the norm, pay the grasshopper and starve or renege on the agreement and watch the grasshopper starve? Here we will have a genuine moral dilemma, in which the ant must choose between justice and its life. The ant may choose death before dishonour, but may also choose to renege with good authority. Thomas Aquinas writes:

> if the need be so manifest and urgent that it is evident that the present need must be remedied by whatever means be at hand (for instance when a person is in some imminent danger, and there is no other possible remedy), then it is lawful for a man to succor his own need by means of another's property, by taking it either openly or secretly: nor is this properly speaking theft or robbery.[7] [2], Question 66, Article 6.

Thus the ant has a choice, and either option can be justified. What the ant will do will depend on its value preferences. Arguably the original contract was foolhardy - on the part of both - since the failure of the harvest could have been foreseen by both parties, and whichever suffers has only themselves to blame.

4 What Makes a Moral Audience?

As the last example shows, there may be more than one morally acceptable ordering on values. Some other orderings, such as a refusal to pay the grasshopper even when there a surplus available to do so, are not acceptable. What we must do is to provide our agents with an acceptable ordering on which to base their reasoning. In order to do so, we need to look at the value order prevailing in society. As noted in work on AI and Law, the decisions made by courts often manifest an ordering on values. The case law decisions often turn on the value preferences the judge wishes to express. This use of social purposes to justify judicial decisions was introduced to AI and Law in [13] and more formally presented in [12]. Thus we may look to the law as one source for our value orderings: the assumption being that the moral order is at least compatible with the order reflected in legal decisions. Note that this legal order need not be static and may reflect changing social views and priorities. Although courts are supposed to be bound by precedents (the doctrine of *stare decisis*) as noted by Mr Justice Marshall in the US Supreme Court case of *Furman v Georgia* (408 U.S. 238 1972) there are occasions when "*stare decisis* would bow to changing values".

Several methods of deriving an audience, in the sense of a value ordering, from a set of cases have been proposed. In AGATHA [18] the value ordering which best explains a set of cases was discovered by forming a theory to explain a set of cases, and then attempting to provide a better theory, in terms of explaining more cases, until the best available theory was found. In [11], given a VAF and a set of arguments and a set of arguments to be accepted, the audiences (if any) to which that set is acceptable is determined by means of a dialogue game. Note that the ordering may not be fully determined (a *specific* audience): it may be possible that the desired set of arguments can be accepted by several audiences, represented as a partial order on the values. In [28], the VAF is rewritten as a meta-level argumentation framework [29], from which value orderings can emerge, or be formed, as a result of dialogue games based on the rewritten frameworks. In this last work explicit arguments for value orderings can be made in the manner of [30].

As well as legal cases, we can identify the approved value orderings from stories, using techniques for deriving character motives from choices with respect to actions, originally targetted at explaining the actions of people involved in legal cases [16]. Stories are often used to persuade people to adopt particular value orders, as with the fable and the parable we have considered in this paper. The notion of using didactic stories as arguments for value orderings was explored in [15] and [14]. Since stories like fables and parables were written specifically to advocate particular value orderings, they are highly suited to our purposes. The values concerned are typically clear, the choices sharp and the correct decisions clearly signposted, leaving little room for doubt as to the recommended preference.

We do not propose data mining or machine learning methods here. Although such methods can discover norms from a set of cases represented as facts and outcomes (e.g [32]), the discovered norms derive their authority from the amount of support in the dataset. They are suited to finding rules, but not exceptions, and it is exceptional cases, where norms need to be violated, that interest us. In law, however, single cases may form an important precedents, identifying apparent exceptions to existing norms, closing gaps and resolving conflicts,

[7] This would, of course, also justify the grasshopper stealing from the ant.

often revealing or choosing between value orderings as they do so.

As noted above, these methods may produce not a specific audience, but a set of audiences all of which conform to and explain the prevailing decisions. If this is so the question arises as to whether it is desirable or undesirable for all agents to be drawn from the same audience. To unify the audience would be to impose the designer's view as to what is moral, albeit constrained by the social decisions. In practice a degree of diversity may prove useful, leading to different agents occupying different social roles.

5 Summary

In this short position paper we have taken as our starting point the idea that as the use of agents speads and as they adopt the autonomous performance of ever more critical tasks, including perhaps, in the not very distant future, warfare and counter terrorism, there is a need to provide them with the capacity for moral reasoning. We have argued that neither of the approaches popular in current multi-agent systems, the enforcement of norms by the removal of the capability of violation, or the provision of a set of guiding principles will enable this. Moral behaviour requires and includes the recognition that on occasion it is right to violate norms, because while norms may be best observed in an ideal world, we need to be able to cope with the sub-ideal, and with the unforeseen. Unforeseen events may occur which mean that following a norm results in underdesirable effects, perhaps even subverting the very values the norm was designed to promote. Moreover when another agent transgresses norms, so producing a sub-ideal situation, it may be necessary to deviate oneself, either to punish the transgression or because the case is altered, and in the particular circumstances two wrongs *do* make a right.

But violation of a norm for moral reasons presupposes that the agent can recognise when the norm should be violated and what form the violation should take. This in turn requires that the agent be able to reason morally from first principles, by which we mean apply an ordering on values to the current situation. If we provide agents with a suitable value ordering, and the capacity to apply this value ordering when selecting an action, we can rely on the agents to make moral choices which might not be the case if they were to blindly follow a fixed set of norms. We have identified work which provides the basis for such a capacity. In doing so we provide a morality in the virtue ethics tradition of Aristole [3], as opposed to the consequentialism and deontology represented by current MAS approaches.

The literature also offers a number of approaches in which the moral orders for various societies can be derived from the legal decisions taken and the stories told in those societies. Note that we would expect both inter and intra cultural variation, and evolution over time.

Such matters can be explored and evaluated through simulations of the sort found in [26] and [27]. For a finer grained, qualitative evaluation, the techniques developed can be applied to classic moral dilemmas such as whether a diabetic may be allowed to steal insulin from another (the Hal and Carla case discussed in [19]) and Phillipa Foot's famous *Trolley Problem* [21].

Future work will need to investigate several aspects of value based reasoning, including: inducing value orderings; consideration of the extent to which values are promoted/demoted; and how value orderings can be applied to situations that differ (in some tangible way that suggests novelty) from the ones that originally gave rise to them.

REFERENCES

[1] Aesop, *Fables, retold by Joseph Jacobs*, volume Vol. XVII, Part 1, The Harvard Classics. New York: P.F. Collier and Son, 1909-14.

[2] Thomas Aquinas, *Summa theologica*, Authentic Media Inc, 2012, written 1265-74.

[3] Aristotle, *The Nicomachean Ethics of Aristotle, translated by W.D. Ross*, Heinemann, 1962, written 350BC.

[4] I. Asimov, *I, Robot*, Robot series, Bantam Books, 1950.

[5] K. Atkinson and T. Bench-Capon, 'Taking the long view: Looking ahead in practical reasoning', in *Proceedings of COMMA 2014*, pp. 109–120.

[6] K. Atkinson and T. Bench-Capon, 'Practical reasoning as presumptive argumentation using action based alternating transition systems', *Artificial Intelligence*, **171**(10), 855–874, (2007).

[7] K. Atkinson and T. Bench-Capon, 'Addressing moral problems through practical reasoning', *Journal of Applied Logic*, **6**(2), 135–151, (2008).

[8] K. Atkinson, T. Bench-Capon, and P. McBurney, 'Computational representation of practical argument', *Synthese*, **152**(2), 157–206, (2006).

[9] T. Bench-Capon, 'Persuasion in practical argument using value-based argumentation frameworks', *Journal of Logic and Computation*, **13**(3), 429–448, (2003).

[10] T. Bench-Capon, 'Transition systems for designing and reasoning about norms', *AI and Law*, **23**(4), 345–366, (2015).

[11] T. Bench-Capon, S. Doutre, and P. Dunne, 'Audiences in argumentation frameworks', *Artificial Intelligence*, **171**(1), 42–71, (2007).

[12] T. Bench-Capon and G. Sartor, 'A model of legal reasoning with cases incorporating theories and values', *Artificial Intelligence*, **150**(1), 97–143, (2003).

[13] D. Berman and C. Hafner, 'Representing teleological structure in case-based legal reasoning: the missing link', in *Proceedings of the 4th ICAIL*, pp. 50–59. ACM, (1993).

[14] F. Bex, K. Atkinson, and T. Bench-Capon, 'Arguments as a new perspective on character motive in stories', *Literary and Linguistic Computing*, **29**(4), 467–487, (2014).

[15] F. Bex and T. Bench-Capon, 'Understanding narratives with argumentation', in *Proceedings of COMMA 2014*, pp. 11–18, (2014).

[16] F. Bex, T. Bench-Capon, and K. Atkinson, 'Did he jump or was he pushed?', *AI and Law*, **17**(2), 79–99, (2009).

[17] D. Bošnački and D. Dams, 'Discrete-time promela and spin', in *Formal Techniques in Real-Time and Fault-Tolerant Systems*, pp. 307–310. Springer, (1998).

[18] A. Chorley and T. Bench-Capon, 'An empirical investigation of reasoning with legal cases through theory construction and application', *AI and Law*, **13**(3-4), 323–371, (2005).

[19] G. Christie, *The notion of an ideal audience in legal argument*, volume 45, Springer Science & Business Media, 2012.

[20] Phan Minh Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial intelligence*, **77**(2), 321–357, (1995).

[21] P. Foot, *Virtues and vices and other essays in moral philosophy*, Cambridge Univ Press, 2002.

[22] R. Hare, *Freedom and reason*, Oxford Paperbacks, 1965.

[23] H. Hart, *The concept of law*, OUP Oxford, 2012.

[24] A. Jones and M. Sergot, 'Deontic logic in the representation of law: Towards a methodology', *AI and Law*, **1**(1), 45–64, (1992).

[25] R. Lipsey and K. Lancaster, 'The general theory of second best', *The review of economic studies*, **24**(1), 11–32, (1956).

[26] M. Lloyd-Kelly, K. Atkinson, and T. Bench-Capon, 'Emotion as an enabler of co-operation.', in *ICAART (2)*, pp. 164–169, (2012).

[27] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, and M. Bench-Capon, T.and Luck, 'Establishing norms with metanorms in distributed computational systems', *AI and Law*, **23**(4), 367–407, (2015).

[28] S. Modgil and T. Bench-Capon, 'Integrating object and meta-level value based argumentation', in *Proceedings of COMMA 2008*, pp. 240–251, (2008).

[29] S. Modgil and T. Bench-Capon, 'Metalevel argumentation', *Journal of Logic and Computation*, 959–1003, (2010).

[30] Sanjay Modgil, 'Reasoning about preferences in argumentation frameworks', *Artificial Intelligence*, **173**(9), 901–934, (2009).

[31] Ch. Perelman, *The new rhetoric*, Springer, 1971.

[32] M. Wardeh, T. Bench-Capon, and F. Coenen, 'Padua: a protocol for argumentation dialogue using association rules', *AI and Law*, **17**(3), 183–215, (2009).

[33] M. Wooldridge and W. van der Hoek, 'On obligations and normative ability: Towards a logical analysis of the social contract', *J. Applied Logic*, **3**(3-4), 396–420, (2005).

A.I. for Online Criminal Complaints: From Natural Dialogues to Structured Scenarios

Floris Bex, Joeri Peters and **Bas Testerink**[1]

Abstract. There exists a mismatch between the sort of crime reports that police would prefer to have and the stories people tell when filing a criminal complaint. Modern crimes such as trade fraud can be reported online, but a combination of static interfaces and a follow-up process that is dependent on manual analysis hamper the intake and investigation process. In this paper, we present our project Intelligence Amplification for Cybercrime (IAC), in which we aim to apply various AI techniques to allow natural dialogues about fraud cases. In this way, different parties such as citizens registering a complaint and police investigators can interface with cases composed of scenarios and evidence through natural language dialogues. This will not only solve an urgent practical problem, but also lead to new insights regarding computational models of evidence assessment.

1 Introduction

Reasoning in police investigations is a complex process, which consists of collecting, organizing and assessing a mass of unstructured and unreliable evidence and scenarios in a case [11]. Artificial Intelligence has proposed various scientifically founded ways of treating evidence using, for example, Bayesian networks [12, 22] or non-monotonic logics [7, 24, 15]. One problem for these A.I. models is that most people involved in the investigative process (e.g. detectives, prosecutors, witnesses) do not have the background to be able to to construct and utilize logical or probabilistic models of a case. Instead, the focus in real cases is often on more linguistically oriented concepts such as *arguments* and *scenarios*, often rendered informally (e.g. natural language) or semi-formally (e.g. mind-maps, argument maps). While recent research has tried to integrate arguments and scenarios with logic and probability theory [29, 23, 28], there still exists a clear gap between real investigations and more formal models [27]. This not only limits the practical applicability of A.I. models, but also makes it very difficult to validate whether formal models are useful and appropriate for investigative and decision-making practices (cf. [26, 25]). What is needed are technologies and theories for the process of investigation that bridge the gap between natural language interfaces and more formal models of evidential reasoning. This paper discusses such technologies and theories in light of a practical application, namely the improvement of online criminal complaints and the subsequent investigation.

Our project Intelligence Amplification for Cybercrime (IAC) aims to develop smart technologies to improve the online intake of criminal complaints and the subsequent investigations on the topic of e-crime and cybercrime. The possibility of reporting a crime online is relatively new in the Dutch police organisation, and it is currently only possible to report a few types of crime. One of these types concerns so-called e-crime, online fraud cases such as fake webshops and malicious second-hand traders. There are about 40,000 complaints about these types of cases every year, and while the damages in each individual case are usually quite small (around 50-100 euros), it pays to follow up on such complaints, particularly because suspects may be part of a larger criminal organization. The high volume and relatively low detail of such cases thus makes them ideal for online complaints and further automated processing.

In this paper, we sketch the outline of a system for reporting e-crime. In its current incarnation, this system consists of a dialogue interface and a module that translates structured and unstructured free text input from the dialogue interface to knowledge graphs, a labelled graph containing the entities, events and relations in a case. Citizens that want to file a complaint can thus tell a story about why they think they were victims of fraud. This story is then automatically translated to a graph that contains the evidence and (possible) scenarios in the case. This graph can then be used for further formal analysis, or to ask questions about the case in the dialogue, further eliciting relevant information from the person who makes the complaint.

The rest of this paper is structured as follows. In Section 2, we describe the application domain of online trade fraud by giving some examples and discussing the current intake process for criminal complaints. In Section 3, we explain the general architecture of our solution. Sections 3.3 and 3.2 delve into the applied computational linguistics and the application of argumentation dialogue literature. We outline in Section 3.4 how the improved structured data can support the police processes that are involved in online fraud. Finally, we conclude with our conclusions and future plans in Section 4.

2 Online Trade Fraud

The Dutch National Police has a number of possibilities for registering a criminal complaint online. Most of the crimes that can be reported online are low-profile, high volume crimes for which there is no clear suspect – for example, bike theft or petty vandalism. However, the National Service Centre E-Crime of the Dutch Police is currently involved in a pilot where citizens can file a complaint for online fraud cases, such as fake webshops and malicious second-hand traders, where there is a clear indication of a suspect (usually because the victim has transferred money to a bank account).

2.1 Typical Trade Fraud Scenarios

As examples of fraud scenarios, consider the types of fraud that take advantage of (first- and) second-hand auction websites, of which

[1] Utrecht University, the Netherlands, email: {F.J.Bex,J.G.T.Peters,B.J.G.Testerink}@uu.nl

ebay.com is probably the most recognised. A similar auction site that is very popular and well-known in the Netherlands is called Marktplaats(.nl) (lit. market place), which we will take as the example. The most obvious type of trading fraud is when swindler Y creates an ad on Marktplaats, advertising a product. Victim X responds to this ad and decides to buy the good. X then transfers the agreed-upon amount of money to the bank account provided by Y, but Y does not send the product (Figure 1). In the case of genuine fraud, the name and bank details provided by Y are most likely false, possibly belonging to a so-called "money mule". This is a person whose bank account is used for criminal activities, wittingly or otherwise.

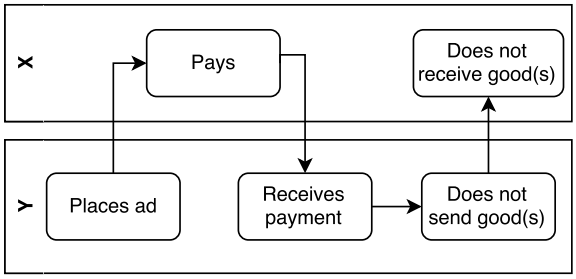


Figure 1. The classic swindle between victim X and swindler Y.

A more elaborate construction can be found in what may be translated as the "triangle swindle" which is depicted in Figure 2. From the point of view of the victim, there is no difference with the first classic scenario. However, the person to whom victim X's payment was transferred is not swindler Y, but rather person Z. Z received X's payment after having been contacted by Y about Z's ad on Marktplaats, after which he sent the goods to the address presented by Y. All this happened because Y copied Z's ad and allowed X to pay for the original ad. Not only has Y received this good for free, X believes that Z is the culprit and Z is not even aware of any complications until he is accused by X.

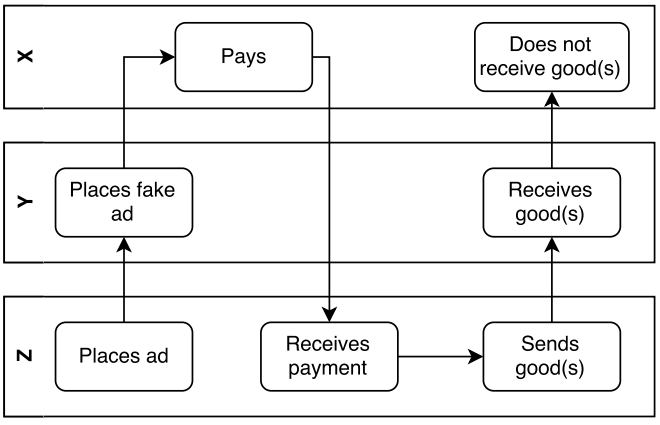


Figure 2. The triangle swindle between victim X, swindler Y and person Z.

Other possible scenarios do not involve trade sites such as Marktplaats, but involve more elaborate spoofed websites. Criminals create websites which seem almost identical to the original web shop they try to imitate. There are signs, of course, such as strange URLs and significantly lower prices. But some people will fall for this, genuinely believing that some big web shop has set up a discount version of their own shop. They are also persuaded by web shop certification marks, even though these are just images freely available on the internet. These spoofed websites allow for large groups of people to be scammed at once, by not sending them their orders. There is a variety of this spoofed website scenario where the web shop is not even trying to imitate another. These websites are completely registered at an address and with the Chamber of Commerce. They can be contacted by telephone and have a registered owner (a money mule). All this attention to detail makes them appear even more trustworthy, so they result in high amounts payments from victims.

2.2 Online intake of Trade Fraud Complaints

The current method of submitting a complaint consists of filling out an online form with some basic information, such as the complainant's details, details about the good or service that the citizen tried to purchase and any available details about email addresses, aliases and bank accounts used by the suspect. Furthermore, the form also has a free-text box in which the complainant is asked to fill in "what happened?" The form's contents are submitted to the police. The complainant is notified and might be contacted at a later stage regarding a follow-up investigation.

At the National Service Centre E-Crime, human analysts further manually analyse those entities (bank accounts, email addresses) from complaints that are suspicious. For example, if a particular bank account pops up in multiple complaints, there might be a fraudster at work. The analysts then take this entity and all related information from the different complaints, and visualize this as a "cluster", a mind-map showing the relations between entities such as bank accounts, aliases, URLs, and so forth and the basis of such clusters, and an accompanying Excel file, the case is built. First, further evidence is gathered from, for example, banks, e-commerce websites and internet service providers. The original complainants in a case are also contacted and asked for more evidence (email conversations with the suspect, screenshots). On the basis of this evidence, one or more scenarios are constructed about exactly what type of fraud has taken place.

One of the problems of the intake and investigation process on trade fraud is that there is a disconnect between the online form that a complainant fills in (i.e. the intake), and the construction and analysis of scenarios based on evidence (i.e. the investigation). The complainant does not always know exactly which information the police or judiciary need to follow up on a case. For a fraud, the victim should have been misled by false contact details, deceptive tricks or an accumulation of lies [2] – these are legal terms for which it is not immediately apparent exactly what they mean or how they should be proved. For example, if a victim was convinced to pay because the suspect offered the lowest price, then this alone is not enough for a fraud case. However, if the suspect also imitated a trusted party, the chances of successfully convicting the suspect for fraud increase significantly.

Because of the various subtleties mentioned above, it is often not clear at the time of filing the complaint exactly what sort of fraud (if any) has been committed. This is less of a problem for the regular intake process: when a complaint is filed at the police station, the complainant can state what happened to a police officer, who can match the incident to known fraud scenarios and ask further questions to try to confirm which particular type of scenario is applicable for the complainant's case. The online form is static and not connected to any possible fraud scenarios (which are only constructed

[2] Article 326 of the Dutch Criminal code

after multiple manual analysis steps), so it is impossible to ask any questions at the time of the intake.

Our initial aim is twofold: first, we need to connect the intake and investigation processes, so that the complainant can describe the incident and the system will directly structure it into scenarios and try to match it to known scenarios. Second, we want to make the intake process more dynamic through dialogue interactions. The subtle details in the scenarios are important but hard to capture and maintain in a large decision tree. So what is needed is a dialogue system that is able to ask the right questions based on the details given by the victim and on the scenarios that might be applicable. There are different fields in artificial intelligence that offer solutions for aspects of this application: multi-agent dialogue systems, computational linguistics and argumentation theory.

3 An A.I. System for Dialogues About Trade Fraud Scenarios

In this section we explain the general architecture of the proposed application for intake and investigation, with a focus on the different solutions from artificial intelligence that we use. Note that the work is ongoing: at the moment, the general agent architecture (Figures 3 and 5), the dialogue interface (Section 3.1) and natural language processing module (Section 3.3) have been developed and connected to each other. For dialogue management (Section 3.2) we will need to tailor the generic framework DGEP by Bex et al. [4]. Scenario reasoning (Section 3.4 is currently limited to basic ontology queries, but will be expanded with more advanced reasoners for argumentation-based semantics in the future.

Our application is a good example of a hybrid system containing both sub-symbolic artificial intelligence for machine learning and language processing, and symbolic artificial intelligence to reason about reports and cases. The system as a whole implements a dialogue system [17], and thus captures the process of intake and investigation. There are two main types of users: complainants who file new criminal complaints, and the police who want to analyse reports and combine them into a case file (Dutch: *proces-verbaal*).

A high-level overview of the system is shown in Figure 3 (for a more detailed view that focuses on the Natural Language Processing System see Figure 5). The complainant and police interact with the system through a dialogue interface. This interface allows users to submit input, i.e. make dialogue moves, but also shows the status of the dialogue such as the open questions. Questions can be generated by both the complainant and the police, but will also originate from the system itself through the scenario reasoning module. The dialogue is managed by a dialogue manager that maintains the legal moves of the participants. The legality of a move for a participant is based on the participants' commitments in the dialogue (e.g. statements that were made). The maintenance of the commitments in a commitment store is also part of the dialogue manager and its details are explained in [4]. The natural language processing system is called upon in case a participant provides free-text input. This system also maintains a knowledge graph that is constructed throughout the dialogue (Section 3.3. The graph serves as input for the scenario reasoning module of the application which then, based on the status of what is known about the reporter's incident, asks extra questions and clarifications through the dialogue manager. Finally, the scenario reasoning module also provides the analysis of reports and cases to the police.

We have opted to design the natural language processing and scenario reasoning components of the application with the agent

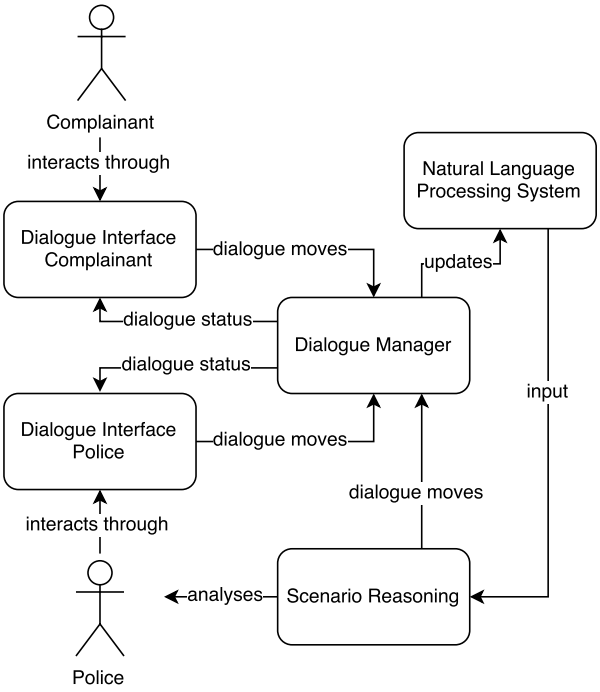


Figure 3. Architecture of the intake system. Boxes indicate software modules. Arrows indicate interaction between components such as service calls or input provision.

paradigm [21]. We use the object-oriented agent programming framework by Dastani and Testerink [10] to implement these components. This program is an object oriented translation of the logic-based programming language 2APL [9]. The agent paradigm matches modules of the software with high-level concepts such as beliefs, knowledge, goals and strategies. One of the main reasons for the agent paradigm is the dialogical nature of intake and investigation. It also benefits maintenance and eases the explanation of the software to outsiders. The agent paradigm also helps with our modularity goals as agents consists of modular components that implement their capabilities. The use of an object-oriented framework, in contrast to many logical approaches in agent oriented programming, not only further supports modularity but is also more accessible to programmers outside of academia. Finally, agent oriented software is distributed in nature, which accommodates the distribution of the application over a cluster of computers. We require such distribution for machine learning purposes, but also because the data is physically distributed and it is easier to attach an agent to data locally rather than collect all the data in a central location.

3.1 Dialogue Interfaces

Both the reporter of a crime as well as the police interact with the application through a dialogue interface. Figure 4 shows the current interface, in which a complainant, Mr. Smith, talks with a police agent, which may be a human or a software agent.

We recognize that natural language processing becomes increasingly more accurate, but also that sometimes it is easer for the application, or more comfortable for the user, to use a (partial) form. The dialogue that we envision will be a combination of free-text and forms, where it is often possible to switch between forms and free-text. For the interface layout we use basic web-based technologies. However, we also include speech-to-text (STT) and text-to-speech

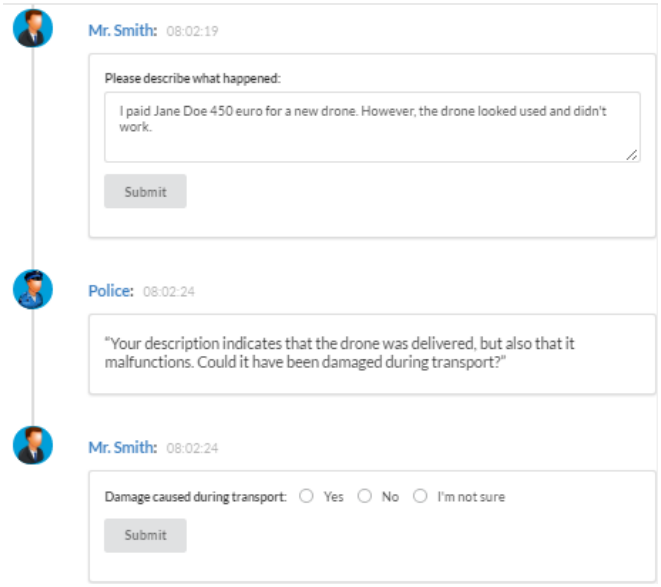


Figure 4. Screenshot of the dialogue interface

(TTS) solutions. These solutions are becoming more widely available and are getting increasingly more accurate. Various companies have now released free-to-use STT and TTS software for mobile devices. As part of future work, we want to investigate whether a keyboard-free, or perhaps even screen-free, dialogue system improves the user experience over a regular web interface. However, we focus on regular input for now as these technologies do not yet approach the accuracy which we feel will support a comfortable experience.

3.2 Dialogue Management

The dialogue that a complainant has with our application is not a completely free-text dialogue, but a mix between text and small forms. Also the topic of the conversation is strongly restricted to the task of reporting a crime which is well defined. This factor greatly reduces the amount of uncertainty that is encountered during the dialogue. Hence, we do not require a statistical approach to dialogue management (e.g. by modelling the dialogue as a Markov Decision process [16]).

The main aim of the system is to provide analyses of reports in order to support cases against swindlers. Because the analysis of scenarios and evidence can be naturally rendered in an argumentation formalism [1], our main approach to dialogue management comes therefore from argumentation dialogue systems theory. We take as a starting point the Dialogue Game Execution Platform (DGEP [4]). This platform allow us to specify the 'rules of the dialogue', such as turn taking, and then provides a mechanisms that keeps track of the commitments of a user. For instance, if a user states that he/she paid a particular seller, then we may commit that user to providing an argument why that particular seller was chosen over other alternatives. We note that not all information can be obtained from the user, hence some commitments are up to the police to satisfy. An example of this is the identifying information of a bank account which is one of the requirements to argue that a swindler used a false identity. The input that the users provide is stored in a knowledge graph that is explained in more detail in Section 3.3.

3.3 Natural Language Processing

In the agent-oriented architecture, there are basically two main types of agents: agents that enhance user input with extra knowledge sources and an agent that reasons about the known facts regarding a report. Both of these agents use a central knowledge base or scenario blackboard. Figure 5 shows a more detailed system architecture. In this section we discuss the agents of the first type (agents that enrich the user's input, i.e., the classifier, ontological, parsing and lexicon agents).

The textual nature of crime reports requires us to address natural language processing: for scenario reasoning, the relevant scenarios and entities from a criminal complaint are needed in a structured form, so that scenarios can be matched to typical fraud scenarios and further reasoned about using the evidence in the case (section 3.4). The task of the natural language processing module is therefore to identify entities in a text (e.g. companies, individuals and products) and then find the expressed relations between them (e.g. person A swindled person B, website C imitates company D).

We focus on (semi-)supervised techniques, where hand-crafted knowledge engineering is part of the design. Knowledge engineering comes in the form of ontology design (based on description logic) to specify the types of entities that we are interested in and the possible relations among them. A scenario model is an ontology plus extra instances of relations and concepts (the identified entities in a specific incident). There exist various frameworks for developing and reasoning with ontologies such as Protégé[3]. The widespread availability of triple storage and query technologies for ontology systems (such as Fuseki[4]) allows us to straightforwardly create a black board where agents can add and retrieve data. While it is challenging in general to define an ontology that covers all the necessary concepts, we have the advantage of having a specific domain: our ontology can be built using domain experts, existing crime reports and judicial documents.

The input from the dialogue interfaces is turned into a directed labeled graph called a *knowledge graph* (Figure 6). The graph combines the different knowledge sources such as the ontology for fraud cases that we use. Hence entities are adopted as nodes in the graph, but also other types of nodes exist, such as the words in the text input in the dialogue interface. The edges represent relations among the nodes, where labels of edges identify the type of relation. The objective of our natural language processing module is to predict new edges among entities in the knowledge graph. We illustrate our approach by assuming we have just the following two sentences.

1. "*I paid John*"
2. "*I paid no attention to the URL*"

Given these sentences we want to predict whether some type of payment has taken place. We do this by predicting whether a pay-edge exists between two identified entities. This is the case in the first sentence where the reporter pays a person named John. The final knowledge graph given the two sentences text is given in Figure 6, which we shall construct throughout the rest of this section.

The dialogue manager initiates the knowledge graph by inserting the words of the user as nodes in the graph. The parser agent then relates these words to each other by using a dependency parser – in our case, the Alpino parser for Dutch [8]. In our example graph, we use outgoing edges only for head dependents (i.e. 'paid', 'attention', 'to' and 'URL' are heads of (sub)dependency trees). A label 'X/Y' between words indicates that the dependency is of type 'X' and the

[3] http://protege.standford.edu
[4] https://jena.apache.org/index.html

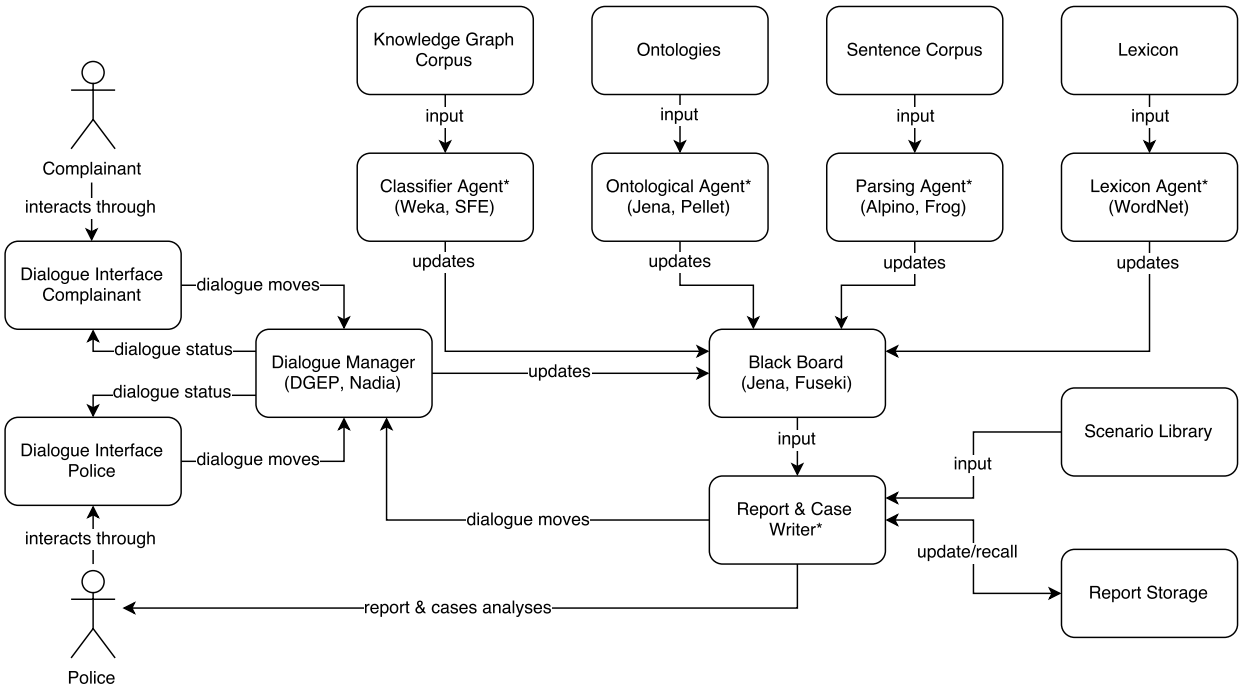


Figure 5. A more detailed system overview of the natural language processing module. Between parenthesis are possible alternatives to support the implementation of a system component. Components with an asterisk are implemented as an agent module and can be instantiated as independent agents or be combined in a single agent.

referred to word is on top of a syntactical tree of type Y. For example, in the second sentence there is an edge from 'paid' to 'to' labeled 'obj2/pp'. This indicates that 'obj2' is the name of the dependency relation and 'to' is on top of a prepositional phrase.

A lexicon agent can further to annotate the words with data from a dictionary in parallel to the parser agent. For instance, WordNet can be used for this [18], or in our case its Dutch variant by Postma et al. [19]. Using dictionaries to annotate words provides some foundational contextualisation and topicalisation of the text. For instance we may adopt the direct hypernym relations from WordNet, which group words in a semantic class. The direct hypernyms of 'pay', 'attention' and 'URL' are 'give', 'basic cognitive process' and 'address', respectively, according to WordNet. We represent this with labels 'X/Y' towards the WordNet node '_wordnet' where X is the relation that we extract ('dh' for direct hypernym) and Y is the word for that relation. Aside from WordNet, we may also use other sources

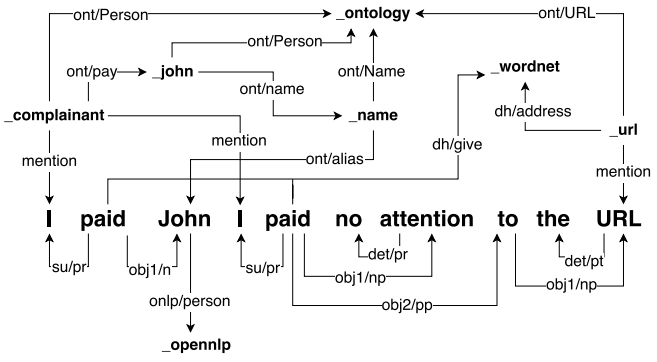


Figure 6. Example knowledge graph.

to classify words such as personal name and organisation name recognition tools. Such a tool would identify 'John' as a person's name, rather than a synonym for toilet as WordNet would classify 'John'. Since we are interested in name recognition, we do make use of this. OpenNLP [5] provides models for finding the names for persons, locations and organisations for both English and Dutch. We adopt the OpenNLP module as a node '_opennlp' and notate the classification of 'John' with the relation 'onlp/person' to indicate that OpenNLP classified 'John' as the name of a person.

The words and relations among those words form the foundation of a special classifier agent that identifies the entities in the sentences. Typically, these are the proper nouns and noun phrases. For the example text we identify the reporter who refers to himself/herself as 'I', we identify an entity named John, and we identify a URL. We adopt these entities as nodes in the knowledge graph, and connect their mentions in the text with a relation named 'mention'. The identification of entities and their mentions throughout a text is a combination of entity resolution and co-reference.

The identified entities in the text are the entities for which we want to determine how they are related to each other and how they are related to implicit other entities (which are not mentioned explicitly in the text). For this we apply classifier agents and ontological agents. Though some of these agents can operate in parallel, most of them have to be iteratively applied in order, because an update of one agent may trigger updates from another agent. Assume we have two disjoint ontological concepts for persons and for URLs. We want to represent in the knowledge graph that '_complainant' and '_john' are instances of persons, and that '_url' is an instance of URL. We do this by inserting the ontology that we use as a node '_ontology' in the knowledge graph, and connect the instances to this ontology by using the ontology's classification as a label edge, that is, '_john' is

[5] https://opennlp.apache.org

connected to '_ontology' with a label 'ont/Person'.

The process of determining the classification of ontological concepts and relations is part of the natural language processing module. We may use a different classifier for each relation that we want to predict. Some of these can be hand crafted. For instance, if an entity has a path with labels 'mention-onlp/person' towards the node '_opennlp', then we may predict that that entity has an 'ont/Person' labeled edge towards the '_ontology' node. Also, if the ontology specifies that a person has a name, then we can insert a name node ('_name') and relate that name to the person entity. Furthermore, a name has an alias. We can assume that an entity that has a name is mentioned by that name, using strings that OpenNLP classifies as person names. For instance, from a '_name' node we may follow all the possible paths ont/name$^-$-mention-onlp/person (superscript hyphen indicates a reverse relation), and then connect the second node of that path (the node that is connected by the mention relation) as an alias to the name. In our example, we identify this way that '_name' has an alias 'John'. Which is the name of the entity '_john'.

However, we may not always have the capability to hand craft classifiers. Furthermore, this may take up a lot of time and is hard to do for a large number of relations. Therefore, we have also looked into possibilities to learn classifiers using machine learning solutions. There are different techniques for learning edge prediction. We have opted for subgraph feature extraction by Gardner et al [13] to determine for an ontological relation what kind of path features (sequences of labels of paths) are predictive of an ontological relation. Consider we want to predict a pay edge between two entities from the text such as '_reporter' and '_john'. A one-side feature is a path feature that does not contain both nodes for which an edge is being predicted. For instance '_reporter' satisfies a path type mention-su/pr$^-$-dh/give. Subgraph feature extraction may find some good features such as '_reporter' satisfies mention-su/pr$^-$-obj/n-mention$^-$-ont/Person (meaning that the entity '_reporter' is mentioned as a subject in a sentence where the object is a person). Given these features and a corpus of example data, we can train classifiers such as support vector machines, or neural networks. We will make use of Weka [14] to actually train the classifiers.

At some point the classifier and ontological agents will signal that they do not have any more information to submit. Then, the agent that interacts with the user will evaluate the current graph and determine whether any further clarification from the user is needed in order to get a complete picture from the incident. The decision about what questions to ask is strongly based on an analysis of what scenarios can currently be constructed. This in turn is also part of the functionality of the scenario reasoning agent that will store the final graph and scenario. Hence these agents overlap quite strongly. If the user is asked for new input, then this input is added to the knowledge graph. Either free text is used, in which case the parsing and lexicon agents have to initiate again, or a form is used, in which case entities and relations can be directly added, and only the classifying agents have to be initiated again.

Finally we note that not all edges and entities can be added through automated means. For instance bank account information or user information from trading websites have to be obtained from third parties. Therefore, there is an interface for the police as well which can be used to add such information to the graph. We expect this information to be added only after the report is filed.

3.4 Scenario Reasoning

Once the information that comes in via the dialogue interfaces is included in the knowledge graph, it will become possible for the scenario reasoning agent to reason with this information. Multiple scenario reasoning agents can participate in a dialogue, using archetypical fraud scenarios from the scenario library and the repository of crime reports. The goal of such agents might be to, for example, match scenarios to typical fraud scenarios, compare scenarios given the available evidence, and elicitate further information from the user. These types of reasoning have been discussed in earlier work on the hybrid theory of scenarios and evidential arguments [6, 7, 1, 28], they have not been implemented. In this section we discuss briefly how implementations of the hybrid theory could be integrated in our system.

Figure 6 represents a (part of a) scenario as a knowledge graph. This knowledge graph can also be rendered as a short scenario, a sequence of events, with supporting evidential arguments, as usually presented in the hybrid theory (Figure 7; note that this figure is not itself a knowledge graph, but rather a more informal rendering of such a knowledge graph aimed at readability). Some elements of the scenario and its supporting arguments can be directly extracted from the knowledge graph: that 'C paid John' and that 'C paid no attention to the URL' can be directly inferred from the knowledge graph based on the complainant's report. The other elements – that 'John built a fake website' and that 'C believed the website was real' will have to be inferred from this (e.g. by applying scenario schemes or by asking the complainant in a dialogue, see below).

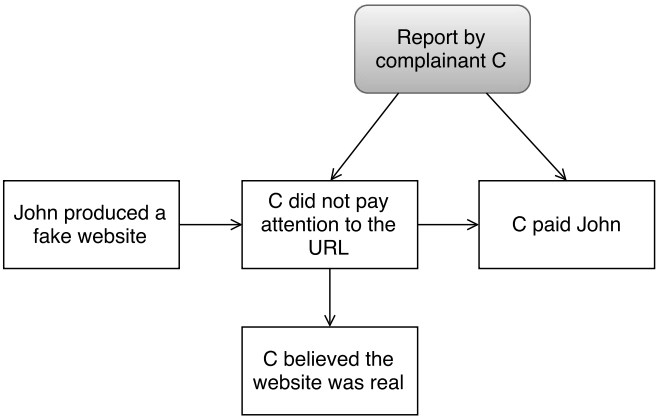


Figure 7. A simple example of a scenario where white boxes represent scenario elements and the grey box an argument (evidence provided by the crime report).

One of the functionalities of the scenario reasoner will be to match the scenario posed by the complainant to typical fraud scenarios known to the police. In other words, it can be checked whether a scenario matches a *scenario scheme* [28]. These scenario schemes can be provided by police experts and are part of the scenario library. Matching knowledge graphs to scenario schemes can be kept relatively straightforward at first. With a few simple rules concerning the presence of entities and relations, specific fraud scenarios can be excluded. Without a mention of a website, for instance, a spoofed website scenario would be a nonsensical.

Excluding scenarios that a complaint does not describe is not necessarily the same thing as identifying the scenario that it does. Often, the complainant cannot distinguish between one scenario and an-

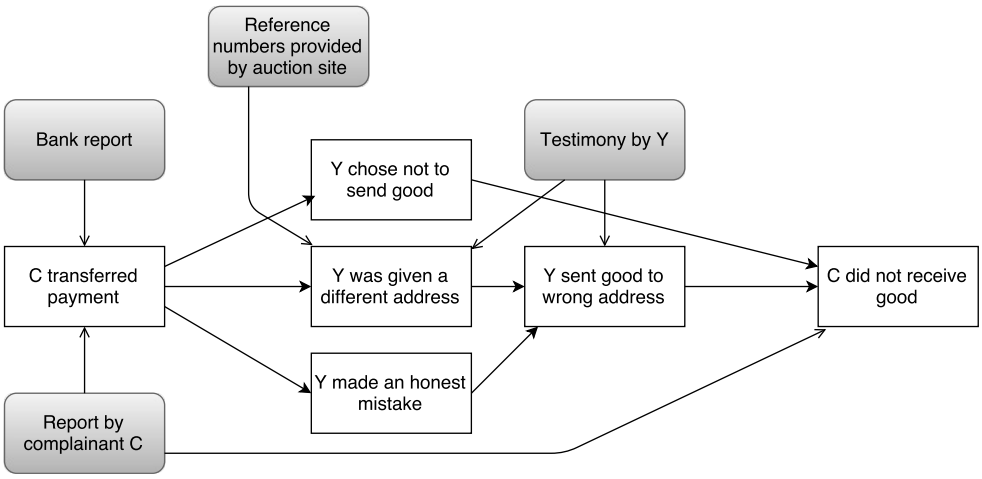


Figure 8. An illustration of one scenario being favoured over another due to an argument provided by the auction site.

other based on their perspective. The triangle swindle, for example, is intended to be indistinguishable from the classic swindle. Depending on the performance of such a rule-based exclusion and keeping in mind the possible future generalisation to other crime types, a more sophisticated solution may be worthwhile. This could be viewed as an exact graph matching problem in a unidirectional graph with labelled nodes and edges.

Determining the exact type of scenario often requires extra evidence. Investigation then becomes a process of inference to the best explanation: there are various possible scenarios that explain why the complainant did not receive the goods, but only one of them is the "true" scenario. Take, for example, the three possible scenarios in Figure 8. The complaint filed online only states that the complainant transferred the payment and did not receive the goods. There are then various explanations for this: the seller Y accidentally sent the goods to the wrong address, or the seller Y chose not to send the goods (a classic scenario, Figure1), or the seller was given a different address by swindler Z (a triangle scenario, Figure 2). Now, if Y testifies that he got the wrong address, and this is backed up by the auction website, chances are that we are dealing with a triangle scenario - of the three possibilities, the triangle option has the most supporting evidence.

The scenario reasoning agent also takes part in the dialogue. During the dialogue, the scenario model can thus lead to a question to be asked. Take, as an example, figure 7. Here, it is not explicitly mentioned by C that he never received the goods - if he would, there would not be a fraud case. From the typical scenario schemes it follows that in a fraud case, the victim should not have received any goods (or the wrong goods). So the scenario reasoning agent can ask the complainant whether they actually received the goods if the complainant did not mention this in first instance. Similarly, the scenario reasoning agent can ask the police analysts for extra evidence. In the situation of Figure 2, for example, the system can ask the police to contact the auction site and Y after the initial complaint, to see whether Y was given the right address. Thus, there are various ways to engage in dialogue about scenarios and arguments [6].

The scenario reasoning agent can ultimately reason with more than just the information from a single complaint. Very often, an investigation incorporates several complaints, it is not uncommon for criminals to be guilty of several types of crime, often even reflecting an overall strategy. The bank account numbers obtained through swindle may be used in another, for example. The hybrid theory allows for reasoning with more and larger cases simultaneously, even if they contradict one another. A combination of crimes is by no means necessarily restricted to trade fraud, as evidenced by the fact that criminals are often identified by linking them with cases from other police divisions. When these cases are themselves linked, such as when a money mule of a type of fraud reports someone for stealing is bank details, this will be reflected in the overall knowledge graph.

For the reasoning about scenarios and arguments to be incorporated into our system, we need to extend our ontology and scenario library with information about arguments - for example, what are the common ways in which a typical scenario or argument can be attacked or extended? Part of this ontology is already captured in the AIF ontology [20], which contains many argumentation schemes and associated critical questions and is available in various common formats (e.g. OWL, RDFs). Another element that the must be captured are the formal semantics of scenarios and arguments [7]. Again, the AIF ontology would be a good fit: as was shown in [5], the status of arguments expressed in the AIF ontology language can be determined using the common argumentation semantics that also underlie the hybrid theory [1].

4 Conclusion

In this paper, we propose a system in which several A.I. techniques are connected. Our system will the police to engage in a dialogue with crime reporters and use the resulting information to their advantage in the subsequent investigation. The system uses sub-symbolic techniques for machine learning and natural language processing to extract a knowledge graph from a complainant's scenario about what happened in a case, and then uses symbolic techniques such as ontologies and argumentative inference to reason with the scenarios and evidence contained in this knowledge graph. Furthermore, the reasoning and processing all takes place in a multi-agent architecture, which allows for modular system development, and is a natural fit with the dialogue interactions and interfaces. Our structured framework of scenarios and evidence establishes the foundation upon which formal reasoning can be applied, and can be used to connect multiple types of police data, which is in line with recent developments surrounding digital filing within the Dutch National Police.

The combination of natural dialogues and structured knowledge graphs will allow us to, for the first time, quickly and relatively simply build and reason about large cases. In the future this will allow for empirical assessment of the various formalisms designed to support evidential reasoning, as the textual dialogue interface allows users with little knowledge of these formalisms to understand and reason about the information in a case. Furthermore, the dialogue interface can also be used for knowledge elicitation. For example, there might be instances in which the classifier cannot accurately predict what type an entity is, or whether there is a relation between entities. The system can then ask a police analyst what the right type or relation should be in that case, and thus extend the ontology. Finally, in the future we also intend to incorporate text generation agents, which will be able to render parts of a knowledge graph as simple textual scenarios.

The techniques developed for our system are generalizable beyond the domain of online trade fraud. The idea of a linked data knowledge graph consisting of scenarios and evidence is applicable to many situations in which the police or judiciary reason with evidence. Two examples are risk assessment surrounding large events [3] and the assessment of asylum applications [2]. Extending the system to other domains will involve a substantial knowledge engineering effort, as scenario libraries will have to be built for different domains (e.g. scenarios surrounding football fan violence [3]. It is further possible to reason with more generic scenarios, such as the 'motivated action' scheme [28] - there are many ontologies available that allow for reasoning with events, time, arguments, and so forth. Finally, we are currently performing data analysis on police data surrounding online crime, which might lead to novel scenarios, frequent patterns that do not correspond to any known scenarios. For example, given the data we have in our project we can try to determine and validate which types of complaints are usually withdrawn (usually because goods have been delivered after all), or designated as being civil rather than criminal (e.g. the delivery of a damaged item or one that is a cheap copy).

ACKNOWLEDGEMENTS

This research is funded by the Dutch National Police Innovation Programme.

References

[1] Floris Bex, 'An integrated theory of causal stories and evidential arguments', in *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pp. 13–22. ACM, (2015).

[2] Floris Bex and Viola Bex-Reimert, 'Evidence assessment in refugee law with stories and arguments', *Informal Logic*, (2016). to appear.

[3] Floris Bex and Bas Hovestad, 'An argumentative-narrative risk assessment model', in *Proceedings of the European Intelligence and Security Informatics Conference (EISIC) 2016*, (2016). to appear.

[4] Floris Bex, John Lawrence, and Chris Reed, 'Generalising argument dialogue with the dialogue game execution platform', 141–152, (2014).

[5] Floris Bex, Sanjay Modgil, Henry Prakken, and Chris Reed, 'On logical reifications of the argument interchange format', *Journal of Logic and Computation*, **23**(5), (2013).

[6] Floris Bex and Henry Prakken, 'Investigating stories in a formal dialogue game', in *Proceedings of the 2008 conference on Computational Models of Argument (COMMA 2008)*, pp. 73–84. IOS Press, (2008).

[7] Floris J. Bex, Peter J. van Koppen, Henry Prakken, and Bart Verheij, 'A hybrid formal theory of arguments, stories and criminal evidence', *Artificial Intelligence and Law*, **18**(2), 123–152, (July 2010).

[8] Gosse Bouma, Gertjan Van Noord, and Robert Malouf, 'Alpino: Wide-coverage computational analysis of dutch', *Language and Computers*, **37**(1), 45–59, (2001).

[9] Mehdi Dastani, '2APL: a practical agent programming language', *Autonomous Agents and Multi-Agent Systems*, **16**, 214–248, (2008).

[10] Mehdi Dastani and Bas Testerink, 'From multi-agent programming to object oriented design patterns', in *Engineering Multi-Agent Systems*, 204–226, Springer International Publishing, (2014).

[11] CJ de Poot, RJ Bokhorst, Peter J van Koppen, and ER Muller, *Recherchеportret - Over dilemma's in de opsporing*, 2004.

[12] Norman Fenton and Martin Neil, *Risk assessment and decision analysis with Bayesian networks*, CRC Press, 2012.

[13] Matt Gardner and Tom Mitchell, 'Efficient and expressive knowledge base completion using subgraph feature extraction', *Proceedings of EMNLP. Association for Computational Linguistics*, **3**, (2015).

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, 'The WEKA Data Mining Software: An Update', *SIGKDD Explor. Newsl.*, **11**(1), 10–18, (November 2009).

[15] John R Josephson, 'On the proof dynamics of inference to the best explanation', *Cardozo L. Rev.*, **22**, 1621, (2000).

[16] Esther Levin, Roberto Pieraccini, and Wieland Eckert, 'Using markov decision process for learning dialogue strategies', in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pp. 201–204. IEEE, (1998).

[17] Peter Mcburney and Simon Parsons, 'Dialogue Games for Agent Argumentation', in *Argumentation in Artificial Intelligence*, eds., Iyad Rahwan and Guillermo Simari, chapter 22, 261–280, Springer, (2009).

[18] George A Miller, 'Wordnet: a lexical database for english', *Communications of the ACM*, **38**(11), 39–41, (1995).

[19] Marten Postma and Piek Vossen, 'Open source dutch wordnet', (2014).

[20] Iyad Rahwan, 'Mass argumentation and the semantic web', *Web Semantics*, **6**(1), 29–37, (feb 2008).

[21] Yoav Shoham, 'Agent-oriented programming', *Artificial intelligence*, **60**(1), 51–92, (1993).

[22] Franco Taroni, Colin Aitken, Paolo Garbolino, and Alex Biedermann, *Bayesian Networks and Probabilistic Inference in Forensic Science*, Wiley, Chichester, 2006.

[23] Sjoerd T Timmer, John-Jules Ch Meyer, Henry Prakken, Silja Renooij, and Bart Verheij, 'Explaining bayesian networks using argumentation', in *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 83–92. Springer, (2015).

[24] Alice Toniolo, Timothy J Norman, Anthony Etuk, Federico Cerutti, Robin Wentao Ouyang, Mani Srivastava, Nir Oren, Timothy Dropps, John A Allen, and Paul Sullivan, 'Supporting reasoning with different types of evidence in intelligence analysis', in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 781–789. International Foundation for Autonomous Agents and Multiagent Systems, (2015).

[25] Alice Toniolo, Timothy J Norman, and Katia P Sycara, 'An empirical study of argumentation schemes in deliberative dialogue', in *Proceedings of ECAI 2012*. IOS Press, (2012).

[26] Susan W van den Braak, *Sensemaking software for crime analysis*, Ph.D. dissertation, Utrecht University, 2010.

[27] PJ van Koppen, *Overtuigend bewijs: Indammen van rechterlijke dwalingen*, Nieuw Amsterdam, Amsterdam, 2011.

[28] Bart Verheij, Floris Bex, Sjoerd Timmer, Charlotte Vlek, John-Jules Meyer, Silja Renooij, and Henry Prakken, 'Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning', *Law, Probability & Risk*, **15**, 35–70, (2016).

[29] Charlotte S Vlek, Henry Prakken, Silja Renooij, and Bart Verheij, 'Building bayesian networks for legal evidence with narratives: a case study evaluation', *Artificial Intelligence and Law*, **22**(4), 375–421, (2014).

The Rise of Smart Justice:
on the Role of AI in the Future of Legal Logistics

Niels Netten and **Susan van den Braak** and **Sunil Choenni** and **Frans Leeuw** [1]

Abstract.
While in business and private settings the disruptive impact of ICT have already been felt, the legal sector is now also starting to face such great disruptions. Innovations have been emerging for some time, affecting the working practices of legal professionals and the functioning and legitimacy of legal systems.
In this paper, we present our vision for enabling the smart government ideal for legal systems by means of a framework that unifies different isolated ICT-based solutions. In particular, we will describe the tremendous potential of improvements driven by AI and challenges to deliver new services that support the objectives of legal systems.

1 INTRODUCTION

In the coming years, the tasks and job descriptions of those involved in the field of law will change dramatically. This change is put into motion by information and communication technology (ICT), which has shown an exponential growth in power, and goes beyond just automating (parts of) current practices [9]. In several domains, ICT already had a disruptive impact on established working practices; thereby creating a completely new industry (e.g., for newspapers and television broadcasters).

In the legal domain, a similar change has also begun as concrete ICT solutions are already emerging to improve and speed up processes. For instance, services performed by computers are replacing the task of document review that has been performed by lawyers up to now. As a result, in big law firms, paralegals, but also high-class lawyers, are being replaced by data scientists. Thus, the expensive and time-consuming process of legal research is being outsourced to a digital expert, who helps with processing massive amounts of relevant legal documents cost-effectively [21].

Another example can be found in civil law in which transactions take place under the auspices of a notary, because they have to take place on the basis of trust and they require to be controlled and registered centrally. In the near future, ICT solutions like a blockchain (a database that represents a public ledger of transactions) can be used as a replacement for such a trusted third party, thereby replacing some of the notary's tasks [23]. Even more drastic changes, [5,21] call them "disruptive technologies", can be expected in the near future as computer-based services are on the verge of replacing other legal tasks: from automatically generating legal documents to predicting outcomes in litigation [14]. Such applications aim to improve specific processes at the operational level. In the end, ICT will affect the working practices of all legal professionals (such as lawyers and judges). The challenge for them will be to take advantage of these developments in ICT and the data generated by ICT-based devices. Otherwise, they (i.e. the lawyers) are outcompeted by others that did apply ICT to innovate [15,21,22] .

Also at the higher tactical and strategic levels, developments in ICT may be exploited, in particular to obtain reliable, valid, and consistent management information. Such tools make use of the increasing amount of data that is generated at the operational level. Management information provides a comprehensive overview of all relevant information needed to make (policy) decisions. It provides insight into the functioning of the system as a whole, and can therefore be used to optimize procedures. This development relates to the smart government vision, which aims to improve government services and to enable collaboration among and participation of government entities, nonprofit agencies, private-sector companies and the public [8,16].

Recent ICT innovations in the legal domain have created many isolated solutions and implementations that generate increasing amounts of data. However, a unifying framework, that exploits these developments and the data produced by them, is missing. Such a framework is needed to streamline innovations and can also be considered as a roadmap towards the smart and open justice ideals. In this paper, we describe such a framework, coined as *Legal Logistics* [17] that utilizes data created by ICT to gain insight and improve processes. This framework unifies separate ICT solutions and enables stakeholders (on various levels) to take advantage of these developments. Subsequently, we will describe our vision on the future of Legal Logistics in view of the developments in research areas such as open data and big data. Moreover, we will give some interesting examples of how Artificial Intelligence (AI) will transform legal systems in the foreseeable future.

2 LEGAL LOGISTICS DEFINED

We define Legal Logistics as a framework *to collect, analyze and integrate all relevant data to gain descriptive, predictive or comparative insight into the functioning of legal systems* [17]. To do so, the objectives of a legal system have to be transformed into meaningful indicators that can be measured using the data available. With this insight the stakeholders (i.e. legal professionals, agencies, researchers and policymakers) can achieve

[1] Research and Documentation Centre, Ministry of Security and Justice, The Hague, The Netherlands, email: {c.p.m.netten; s.w.van.den.braak, r.choenni, f.leeuw}@minvenj.nl.

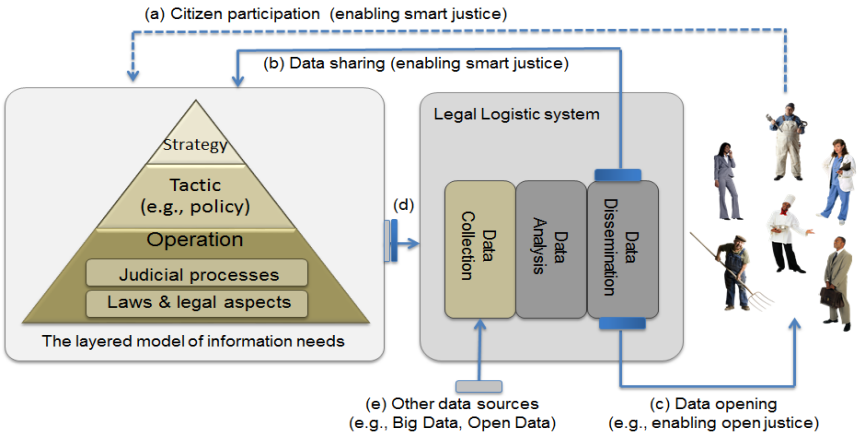


Figure 1: An illustration of the Legal Logistics framework

key objectives and innovate services. In a sense, an information system based on the Legal Logistics framework is a measurement tool to gain insight into the level of well-being of a legal system using statistical information on the past, present, or future state of the system.

Figure 1 shows the Legal Logistics framework. The framework consists of 1) the Legal Logistics system (Figure 1, right rectangle) and 2) various stakeholders who require insight into the legal system. These stakeholders are in turn divided into two groups: 1) legal professionals with different tasks and information needs (Figure 1, left rectangle) and 2) the general public (Figure 1, on the right). The Legal Logistics system represents the technical part of the framework where various ICT solutions are implemented and unified. This system is divided into three stages: 1) data collection, 2) data analysis, and 3) data dissemination.

In the data collection stage, relevant data are obtained from the information systems of the agencies within the legal system (see arrow-d in Figure 1) or from other (external) sources (e.g., social media data, open data, and big data; see arrow-e in Figure 1). In the data analysis stage, the collected data are subsequently exploited to determine relevant and meaningful indicators. In the data dissemination stage, (statistical) information relating to these indicators is shared with agencies within the legal system (see arrow-b in Figure 1) or is disseminated to the public (see arrow-c in Figure 1). In a judicial setting, data dissemination is a separate stage that incorporates procedures to deal with transparency and privacy purposes when sharing the information [2].

With respect to the stakeholders in the framework, the agencies in legal systems have different types of tasks and, as a result, different information needs. These tasks and information needs can be viewed as different layers (see Figure 1, the triangle on the left). The lowest level encompasses operational tasks where legal professionals handle legal cases and enforce laws. Professionals performing these tasks need to share data to perform routine day-to-day operations, for example, to determine the identity of a person. For such tasks, detailed individual-level data often needs to be shared. The middle layer includes the tactical tasks that policymakers or advisers carry out to propose (changes in) policies or legislation. The highest level is concerned with the strategic tasks carried out by top management (i.e., the parliament or ministers): setting long-term objectives and strategies for achieving these objectives For both tactical and strategic tasks, stakeholders need statistical data (i.e., management information) to make informed decisions. Such decisions may involve optimizing tasks on an operational level.

In general, the Legal Logistics framework can be applied to all legal systems that involve these three layers to some extent. It is particularly useful for providing insight into the functioning and performance of a legal system and determining whether its objectives are met.

3 THE CURRENT STATE OF LEGAL LOGISTICS

Within the Research and Documentation Centre of the Dutch Ministry of Security and Justice we have realized a prototype of the Legal Logistics framework for the Dutch criminal justice system.. Our implementation of the framework currently focusses on two purposes 1) providing reliable management information on a tactical or strategic level to improve processes [8,16] (arrow b in Figure 1) and 2) sharing data publicly in order to provide transparency and stimulate the open justice and open government initiatives [1,2, 3,11] (arrow c in Figure 1).

The first purpose is important, because to understand how the criminal justice system functions, it is not sufficient to have access to the information of the separate agencies. Instead, information on the relations and flows between agencies is required. Such insights can only be gained by relating or integrating the data from the different agencies in a coherent manner. However, while the agencies work closely together in the Dutch criminal justice system, this is not as straightforward as it seems [6,16,17].

The second purpose is in line with the vision of a smart government. This aims at using ICT for open-government, open innovation in public agencies, and maximum interoperability among public agencies [13]. A smart government seeks for the best way to serve citizens and the society and aims to improve government services (i.e., by making them quick, measurable, affordable, and sustainable) and enable collaboration among government entities, nonprofit agencies, private-sector companies, and the public [11].

The implemented Legal Logistics framework [17] unifies several information systems [8,16,20] that systematically collect, analyze and disseminate data about the Dutch criminal justice system. These systems mainly concentrate on generating management information by bringing together data coming from different sources. At our research centre, we developed three different systems that are currently used by policymakers and advisers in the Dutch criminal justice system. Each system fulfills a different information need and has a different purpose. More specifically, we developed prototype systems to 1) monitor (case) flows within and between organizations [8] 2) measure elapsed times [16], and 3) predict future workloads [20]).

These systems mainly take the data produced and registered by the agencies within the criminal justice system (such as the police, prosecution, and courts) as an input (arrow d in Figure 1). These data pertain to the criminal cases being worked on and the corresponding offenders. Typically, the collected data sets are highly structured and contain some attributes to identify the case and the suspect/convict involved and some relevant events (temporal data) relating to the procedure followed (e.g., the date the case was received by the agency, the date of important decisions, and the date the case was finished). Our research centre developed a data warehouse and a dataspace system [6,8] for collecting, storing, cleaning, selecting, and integrating the data required. The data are subsequently analyzed and enriched, for instance, by calculating meaningful indicators (such as elapsed time, output, stock, and production). These indicators measure two objectives of a criminal justice system: effectiveness and efficiency [17].

The thus generated management information is shared with governmental agencies, policymakers and advisors in the criminal justice system to allow them to define the future research agenda, to answer policy-related questions, and to assess the effectivity of standing policies. Using this information, they are, for instance, able to detect bottlenecks, unexpected differences, and potential problems, and therefore, they can take better and informed tactical and strategic decisions. As part of the smart government ideal with its open data component, the collected data and enhanced information are also shared with various other external user groups (like scientists, journalists, and the public). In this way, two other objectives of a criminal justice system are met: accountability and transparency. In [17] these objectives are described in detail.

Thus, the implemented Legal Logistics framework shows how data generated by ICT in the Dutch criminal justice system can be utilized to gain and provide insight into the functioning of the system. It provides concrete solutions for collecting and integrating data from various sources, measuring meaningful indicators from these data, and sharing the results. However, this implementation comes with two types of challenges: 1) data quality and semantic interoperability issues when collecting and integrating data and 2) privacy-related issues when disseminating data. Moreover, it lacks a solid feedback mechanism directly to the professionals working at the operational level, since at the moment only statistical management information is being shared. The participation of the public is also open to further improvement. [17] describes in detail how these challenges are to be addressed. Here, we will largely focus on the issues relating to data quality, as AI can play an important role in overcoming some of them.

Agencies in the criminal justice system use administrative information systems with highly structured data to register their operational activities. Since this is not their core business, key problems with the data like, incompleteness, inaccuracy, and inconsistency are not uncommon [6,8,16]. Although there are sound and effective techniques to deal with functional dependencies in the field of data management [4], the management of quantitative (e.g., similar attributes having the same values) and qualitative dependencies (e.g., attributes do not usually show large deviations over time) is mainly left to domain experts and cannot yet be automated fully [6].

External data sources (arrow e in Figure 1) can be a valuable source of knowledge in order to make the data more reliable and complete. In addition, recent developments in AI may be explored in order to represent domain knowledge and automatically handle incompleteness and inconsistencies. How this changes legal logistics in the future will be explained in the next section.

4 THE FUTURE OF LEGAL LOGISTICS

Only recently, as explained in the previous section, the first steps were taken towards connecting and integrating data of multiple agencies resulting in concrete solutions for generating management information for policymakers and advisers on a tactical or strategic level. Given the rapid developments in AI-driven ICT, we envision a future in which all kinds of different data (legal and non-legal; structured and unstructured) are combined and used by smart devices for various purposes and at different organizational levels. Consequently, this will have a tremendous impact on public participation in legal systems, the working practices of the professionals involved, and the nature and validity of the data available for these tasks. This will be explained in the remainder of this section.

In the near future, the main developments in Legal Logistic are to be found in the fields of big and open data. Open data relates to the smart government ideal, which has an open data component. In the foreseeable future we see developments in the direction of semi-open data in order to frame, acknowledge, and encourage such open data initiatives [2,3]. The prospects of big data for the legal domain are very diverse; it could be the key resource for innovative programs and services. In fact, we already see new applications emerging (on the operational level) that use big data techniques such as predictive policing [18] and (text) analysis of criminal files [19].

Big and open data can be exploited at the tactical and strategic level of a legal system for generating more reliable management information. As explained above, currently, only structured administrative or survey data are available and often the quality of these data and the reliability of the sources is uncertain. Additional (big) data sources with semi-structured or unstructured data could be used to analyze the validity of the registered data and complete them. For instance, when information on certain verdicts is missing (or not registered correctly in the information system), this information can be obtained by automatically extracting it from the court files (semi-structured text documents) using text analysis techniques. Furthermore, as another example, social media data may be used to validate survey data obtained through questionnaires, for example, about people's sense of security in public places. Often what people answer in a questionnaire is different to what they actually experienced, for instance, because they do not remember what happened or do not want to answer the question. Moreover, it is hard to determine whether a respondent gives a socially accepted answer or not. Social media are platforms where people usually immediately express what they experienced. Therefore, social media data can, with some caution, be used as an addition to data taken from traditional questionnaires. Another approach to involving citizens in the data collection process, is through crowdsourcing [7].

Another potential use for big data on a tactical or strategic level can be found in the field of policy evaluation. The evaluation of policies is a hard and time-consuming task. Ex-post evaluation studies require, among other things, that baseline measurements, corresponding to the conditions before the start of an intervention, are compared to the observed outcomes at the time the intervention is completed. Usually, between the initiation and completion of an

intervention, there is a long time interval in which many phenomena may occur in the society that also affect the intervention. Consequently, the observed differences cannot be attributed fully to the intervention itself. Using big data may help in better identifying the (hidden and) arising social phenomena in the course of a policy intervention, and compensating their impact on baseline and current measurements.

Beyond these applications of ICT and AI for the foreseeable future, in the more distant future we envision much more disruptive developments. In our vision, the next frontier for legal systems will be cutting-edge AI-driven technology. Such technology will enable advanced, smart, and knowledge-based data analysis, for instance, to determine the context of the data, represent and exploit domain knowledge, and reason with uncertainty. With such techniques, it is, for instance, possible to automate the data integration process mentioned in the previous section, which nowadays requires manual human effort as dependencies need to be handled and domain knowledge needs to be modeled. As the law still relies heavily on reasoning by people, once AI is capable of proper legal reasoning, much will change. To do so, AI-driven applications require adequate domain knowledge and the ability to quickly adapt to changes (in the law or society).

Technologies from the fields of big data and AI are currently already available to help lawyers to prepare for litigation [12]. A digital legal expert called ROSS helps lawyers to "power through" their legal research and aims to keep them up to date on new court decisions that could impact their own ongoing cases. A lawyer can ask questions like "can a company gone bankrupt still conduct its business?" and ROSS gives a cited answer and topical readings from legislation, case law, and other secondary sources after having red through the entire body of law. Similar programs may also support judges with forming a thought-out judgement about the case based on all data and knowledge available. Smart devices can also help to visualize the information to legal professionals in a new manner using augmented reality technology, such as, for example, the HoloLens [10]. With the HoloLens a HoloTrial could be created, a nontraditional method of presenting evidence and legal theories during a trial.

Not only legal professionals will benefit from these developments, it will also be beneficial to citizens and laymen. In our view, the public will have access to (legal) data via personal devices, while these devices will also be able to reason about these data and use domain knowledge. This will help citizens to better understand complex systems, such as the law, and will support them in reasoning about their particular needs (e.g., information about a legal case). As a result, people will be able to participate much more in the legal system. For example, when someone has a dispute with his neighbor about his garden fence, he could consult his legal app on his smart phone to determine which legal steps to take, or whether a case would be successful when brought to court.

However, before AI will play an important role in the future of Legal Logistics, some difficult challenges remain to be overcome. Although the AI domain has made significant gains in learning and decision making under uncertainty, it still faces specific challenges concerning the legal domain. These include, amongst others, the challenge of incomplete knowledge about the world, reasoning with uncertainty and also adapting to a (fast) changing environment and context (e.g., due to changes in legislation) . Therefore, to enable AI-driven technology that will further transform the legal domain, addressing these challenges is becoming more pressing than ever before.

REFERENCES

[1] M.S. Bargh, S. Choenni and R. Meijer, 2016, "Meeting Open Data Halfway: On Semi-Open Data Paradigm", Icegov'16, Montevideo Uruguay.

[2] M.S. Bargh, S. Choenni and R. Meijer, 2016, "On design and deployment of two privacy-preserving procedures for judicial-data dissemination", Government Information Quarterly (GIQ).

[3] M.S. Bargh, S. Choenni & R. Meijer, 2015, "Privacy and information sharing in a judicial setting: A Wicked Problem" In Proc. of the 6th Annual International Conference on Digital Government Research (dg.o), May 27-30, Phoenix, Arizona, USA.

[4] M.S. Bargh, J. van Dijk & S. Choenni, "Dynamic data quality management using issue tracking systems," In the IADIS International Journal on Computer Science and Information Systems (IJCSIS, ISSN: 1646-3692), ed.

[5] Bower, J. L., and C. M. Christensen. Disruptive Technologies: Catching the Wave." *Harvard Business Review* 73, no. 1 (January–February 1995): 43–53.

[6] S. van den Braak, S. Choenni & S. Verwer 2013. Combining and analyzing judicial databases. In Discrimination and Privacy in the Information Society (pp. 191-206). Springer Berlin Heidelberg.

[7] S. Choenni, M. S. Bargh, C. Roepan, and R. Meijer, Privacy and Security in Smart Data Collection by Citizens, in Smarter as the New Urban Agenda: A Comprehensive View of the 21st Century City, Gil Garcia, J. Pardo, T., Nam, T., (eds), Springer, NY, USA, pp.349-367.

[8] J.J. van Dijk, S.N. Kalidien, & S. Choenni 2015 *Smart Monitoring of the Criminal Justice System*. Government Information Quaterly special issue "Beyond 2015 Smart Governance, Smart Development".

[9] M. Fabri and F. Contini, 2001,Justice and technology in Europe: How ICT is changing the judicial business.Kluwer Law International.

[10] HoloLens, 2016, https://www.microsoft.com/microsoft-hololens/en-us Accessed: 1 Jun 2016

[11] R. Howard, 2013 "Smart government key initiative overview", [Online]:https://www.gartner.com/doc/2520516/smart-government-key-initiative-overview.

[12] IBM, 2016, ROSS - Super Intelligent Attorney. URL: http://www.rossintelligence.com/ Accessed: 1 Jun 2016

[13] C. E. Jiménez, 2014. e-Government Interoperability: Linking Open and Smart Government. Computer. 47,10, 22-24.

[14] D.M, Katz, M.J Bommarito and Blackman, J. (2014) Predicting the Behavior of the Supreme Court of the United States: A General Approach. SSRN.

[15] J. O., McGinnis and R. G. Pearce, 2014 The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services. 82 Fordham Law Review 3041.

[16] N. Netten, S.W van den Braak, S. Choenni, S. and Leertouwer E.C. 2014, Elapsed Times in Criminal Justice Systems. Icegov'14, Guimares, Portugal.

[17] N. Netten, Mortaza S. Bargh, S.W van den Braak, S. Choenni, & F. Leeuw. 2016, On Enabling Smart Government: A Legal Logistics Framework for Future Criminal Justice Systems. (dg.o '16), June 08-10, 2016, Shanghai, China.

[18] PredPol, 2016, http://www.predpol.com/ Accessed: 1 Jun 2016

[19] SemLab, 2016, https://www.semlab.nl/portfolio-item/wbom-personalia-extractie/ Accessed: 1 Jun 2016

[20] P. R. Smit and S. Choenni, 2014, On the interaction between forecasts and policy decisions. Dg.o pp:110-117.

[21] R. Susskind, 2008, The end of lawyers? Rethinking the nature of legal services, Oxford: Oxford University Press.

[22] R. Susskind and S. Susskind. The Future of the Professions. Oxford

[23] M. Swan, 2015 Blockchain: Blueprint for a new Economy. O'Reilly Media.

On how AI & law can help autonomous systems obey the law: a position paper

Henry Prakken[1]

Abstract. In this position paper I discuss to what extent current and past AI & law research is relevant for research on autonomous intelligent systems that exhibit legally relevant behaviour. After a brief review of the history of AI & law, I will compare the problems faced by autonomous intelligent systems with the problems faced by lawyers in traditional legal settings. This should give insights into the extent to which AI & law models of legal problem solving and decision support can be applied in the design of legally well-behaving autonomous systems.

1 Introduction

Increasingly, computer systems are being employed in practice with some degree of autonomy. Their behaviour is not fully specified by the programmer but is the result of the implementation of more general cognitive or physical abilities. Such artificially intelligent software can do things that, when done by humans, are regulated by law. To give some example, self-driving cars have to obey the traffic laws. Online information systems that decide whether a system of person can be given access to privacy-sensitive data have to comply with data protection law. Actions of care robots that help sick or elderly people can damage property or the health of the person (spilling coffee over an iPad, failing to administer medication on time). Intelligent fridges that can order food or drinks when the supplies run out have to obey contract law. Autonomous robot weapons have to comply with the of war, with its three principles that soldiers should distinguish between the civilian population and combatants, that an attack is prohibited if the expected civilian harm is disproportional to the expected military benefits, and that military force must be necessary to the explicit purpose of defeating an adversary.

When such autonomous systems are being used, legal rules cannot any more be regarded as regulating human behaviour, since it is not the humans but the machines who act. This raises the problem of how the autonomous systems can be designed in such a way that their behaviour complies with the law. Note that this question needs to be asked irrespective of the question whether machines can assigned responsibility in a legal sense. Even if a human remains legally responsible or liable for the actions of the machine, the human faces the problem of ensuring that the machine behaves in such a way that the responsible human complies with the law.

One solution to the problem is to design the system in a way that guarantees that the system will not exhibit unwanted behaviour. This is the conventional solution when non-autonomous machines, tools or systems are used. [16] called this *regimentation*. A similar approach has been proposed for autonomous systems, such as in the *Responsible Intelligent Systems* project at Utrecht University, which proposes to verify the behaviour of systems off-line with so-called model-checking techniques[2]. However, when systems are increasingly autonomous, their input and behaviour cannot be fully predicted, so that regimentation or advance off-line testing are impossible or of limited value. How can we then ensure that autonomous systems comply with the law? This position paper discusses to what extent the fruits of AI & law research are relevant for solving this problem. (For a related discussion from a more legal perspective and specifically for robots see [18]). To this end, I will first briefly review the history of AI & law research and then compare the problems faced by autonomous intelligent systems with the problems faced by lawyers in traditional legal settings.

2 A brief history of AI & law

The 1970s and 1980s were the heydays of research on knowledge-based systems, such as the influential MYCIN system for diagnosis and treatment of infection diseases [6]). For long[3] computer scientist could in these days easily think that in the legal domain knowledge-based systems can be much easier developed than in the medical and similar domains. While medical knowledge needs to be acquired from human medical experts who are not always aware how they solve a medical problem, legal knowledge would simply be available as rules in written texts, such as statutes and case law reports. And such rules can easily be represented in a rule-based system like MYCIN, after which their application to the facts of a case would be a simple matter of logic. On this account, once a legal text and a body of facts have been clearly represented in a logical language, the valid inferences are determined by the meaning of the representations and so techniques of automated deduction apply.

However, this mechanical approach leaves out most of what is important in legal reasoning, as every lawyer knows. To start with, legislators can never fully predict in which circumstances the law has to be applied, so legislation has to be formulated in general and abstract terms, such as 'duty of care', 'misuse of trade secrets' or 'intent', and qualified with general exception categories, such as 'self defence', 'force majeure' or 'unreasonable'. Such concepts and exceptions must be interpreted in concrete cases, a process which creates room for doubt and disagreement. This is reinforced by the fact that legal cases often involve conflicting interests of opposing parties. The prosecution in a criminal case wants the accused convicted while the accused wants to be acquitted. The plaintiff in a civil law

[1] Department of Information and Computing Sciences, Utrecht University and Faculty of Law, University of Groningen, The Netherlands, email: H.Prakken@uu.nl

[2] https://www.projects.science.uu.nl/reins/, accessed June 2, 2016.

[3] Some parts of his section are adapted from [22] and [23].

suit wants to be awarded compensation for damages, while the defendant wants to avoid having to pay. The tax authority in a tax case wants to receive as much tax as possible, while the taxpayer wants to pay as little as possible. Both aspects of the law, i.e., the tension between the general terms of the law and the particulars of a case, and the adversarial nature of legal procedures, cause legal reasoning to go beyond the meaning of the legal rules. It involves appeals to precedent, principle, policy and purpose, as well as the consideration of reasons for and against drawing conclusions. Another problem is that the law often gives considerable freedom of judgement to the judge, for example, when determining the extent of financial compensation for a tort or when determining the sentence in a criminal case. Although judges are supposed to decide like cases alike, in these matters there are no clear rules, since cases are never fully alike. In all this, it is relevant that the law is not just a conceptual or axiomatic system but has social objectives and social effects, which must be taken into account when applying the law. A final problem is that determining the facts of a case is often hard, since it requires vast amounts of commonsense knowledge of the world, and giving the computer common sense is a recognised hard problem in AI [7].

In sum, law application is not just logically applying statute and case law rules to the facts of a case but also involves common sense, empathy and a sense of justice and fairness. Modelling these aspects in a computer program has so far proved too hard.

However, this does not mean that AI cannot be usefully applied to the law. Deductive techniques have been practically successful, especially in the application of knowledge-based systems in large-scale processing of administrative law, such as social benefit law and tax law. Such systems apply computational representations of legislation to the facts as interpreted by the human user. The use of such systems has been proved to greatly reduce two major sources of errors in the processing of social benefit applications by 'street-level bureaucrats': their incomplete knowledge of the relevant regulations and their inability to handle the often complex structure of the regulations, with complex boolean combinations of conditions, numerical calculations and cross-references [14, 29]. The computer is, of course, ideally suited for retrieving stored information and for handling syntactic and numerical complexities. Deductive rule-based systems have therefore been applied in public administration on a considerable scale. Such systems leave it to the user (the official with the authority to make a decision) to decide whether to accept the system's recommendation or to deviate from it on non-statutory grounds. Thus these systems do not automate legal judgement but the logic of regulations [15, 14].

The deductive model of legal reasoning has been refined with means to express rule-exception structures and hierarchies of regulations. Two common structural features of legal regulations are the separation of general rules and exceptions, and the use of hierarchies over legislative sources to resolve conflicts between different regulations. AI and law has dealt with these features with so-called non-monotonic logics. Such logics have been shown useful in modeling legislative rule-exception structures and legislative hierarchies [10, 24, 13, 32], and in modeling legal presumptions and notions of burdens of proof [25, 11, 12]. Nevertheless, although nonmonotonic techniques technically deviate from deductive logic, their spirit is still the same, namely, of deriving consequences from clear and unambiguous representations of legal rules, rule priorities and facts. More often, conflicts arise not from competing norms but from the variety of ways in which they can be interpreted. A real challenge for deductive accounts of legal reasoning is the gap between the general legal language and the particulars of a case. Because of this gap, disagreement can arise, and it will arise because of the conflicts of interests between the parties.

These observations can be illustrated with the famous *Riggs v. Palmer* case discussed in [9], in which a grandson had killed his grandfather and then claimed his share in the inheritance. According to the applicable inheritance law, the grandson was entitled to his share, but every lawyer understands that he killed his grandfather is a reason not to apply this law. And indeed the court denied the grandson his claim on the grounds that nobody should profit from his own wrongdoing. A deductive or nonmonotonic rule-based system cannot recognise this, unless the exception is already represented in concrete terms in the knowledge base. Adding an explicit exception like 'unless the heir would profit from his own wrongdoing by inheriting' to the relevant legal rules would not solve the problem, since the system cannot recognize that inheriting from one's grandfather after killing amounts to profiting from one's wrongdoing, unless this is explicitly represented in the system's rule base.

Nevertheless, AI offers more to the law than systems based on deductive or nonmonotonic logic. To start with, when for an interpretation problem the relevant factors are known, and a large body of decided cases is available, and these cases are by and large consistently decided, then techniques from machine learning and datamining can be used to let the computer recognize patterns in the decision and to use these patterns to predict decisions in new cases. One example is [8]'s statistical model for predicting whether a job offered to an unemployed is 'suitable employment', in which case refusal of the job offer should lead to a reduction of the employment benefit (see [31] for a neural-network application to the same data and [2] for a similar application to UK social security law). Another example is the sentencing system of [20], which could give judges deciding on sentences for street robberies insight into sentences assigned in similar past cases. On sentencing see also [28].

In spite of the good level of performance of such AI techniques, their practical usefulness in the legal domain is limited, for two main reasons. First, not many legal interpretation problems meet all three requirements for successful use of these techniques: a known and stable set of relevant factors, many decided cases, and little noise among or inconsistency between these cases. More importantly, these techniques are notoriously bad in explaining their output. They are essentially black boxes, which give no insight into how they relate their input to their output. Needless to say that for judges this is a major obstacle to using these systems.

These limitations are addressed in AI & law research on legal argument. This research has led to many important theoretical advances, all based on the idea that legal reasoning is about constructing and critically evaluating arguments for and against alternative solutions of a case. Detailed models have been provided of the role of cases, principles, values and purpose in legal reasoning, of analogical reasoning of reasoning about evidence and of the role of procedure and burden of proof in legal reasoning. For overviews see e.g. [27, 4, 26, 23]. While some of this research has been purely theoretically motivated, others ultimately have practical aims. For instance, [1] sketched a vision of a system which could support an advocate charged with preparing a case at short notice. The system would be able to accept the facts of the case and then generate arguments for the two sides to the case and counterarguments to them, together with the precedents on which they are based. However, such a system is not yet in practical use at any law firm. A main problem with AI & law's proof-of-concept systems has so far that they are critically dependent on the possibility of acquiring a large amount of knowledge and representing it in a form which can be manipulated by the

system. This is an instance of the well known 'knowledge acquisition bottleneck', which has proved a major barrier to the practical exploitation of intelligent techniques in many domains.

The most recent development in AI & law research is a revitalisation of research on information retrieval by the recent spectacular developments in such areas as deep learning, data science, and natural-language processing, combined with the availability of huge amounts of unstructured legal information on the internet. This has put new topics such as information integration, text mining and argument mining on the research agenda. With IBM's Watson system available, the holy grail for many in legal informatics is not an argumentation assistant as described in [1] but a legal research assistant in the form of an application of Watson, which can efficiently find, summarise and integrate information relevant to a case.

Nevertheless, there is some hope that this recent research can also make an argumentation assistant within research. A very recent application of text mining called 'argument mining' has become popular [21, 33, 17] and IBM's Watson team has already experimented with a 'debater' function, which can find arguments for and against a given claim. The fruits of this research can perhaps be combined with AI & law's argumentation models in such a way that these models can finally be scaled up to realistic size, without the need for formal knowledge representation.

3 Is obeying the law always desirable?

Before discussing how autonomous systems can be made to obey the law, first another question must be discussed: it it always desirable to obey the law? In part this is still a legal question, since (parts of) legal systems have general exception categories like the exception concerning self-defence and other ones in criminal law, a general exception in Dutch civil law that statutory rules concerning creditor-debt relations shall not be applied if such application is unreasonable, and so on. Consider the case of the autonomously driving Google car, which was stopped by the California police for driving too slowly. Google had for safety reasons set the car's maximum speed for roads in a 35mph zone at 25mph and one of its cars was causing a big queue of traffic while driving 24mph.[4] From a technical legal point of view this is not a case of undesirable norm obedience, since the relevant traffic regulation contains the following general exception clause:

> No person shall drive upon a highway at such a slow speed as to impede or block the normal and reasonable movement of traffic, unless the reduced speed is necessary for safe operation, because of a grade, or in compliance with law.

However, there is still a practical problem, since general exception clauses like these introduce vagueness and uncertainty. Human drivers are generally good at determining when their speed is to slow by applying their experience and common sense. However, can autonomous cars be given the same kind of common sense? For a preliminary proposal see [19].

One step further are cases in which behaviour is from a technical legal point of view illegal but still socially acceptable. For example, slightly speeding in a queue of cars that all drive a few miles above the maximum speed; waiting for a red pedestrian crossing light at night with no traffic within eyesight; admitting a student to a university course who missed the strict admission deadline for some stupid reason. Here the reasoning problem is logically the same as with general exception clauses: determining whether particular behaviour satisfies some general exception category to a behavioural rule. That the exception is now for social instead of legal acceptability is irrelevant for the kind of reasoning involved.

This all means that the behaviour of autonomous systems should not be seen as rule-governed but as rule-guided. Legal rules are just one factor influencing socially optimal or permissible behaviour. Other factors are e.g. social conventions, individual or social goals or simply common sense. And sometimes these other factors override the legal factors. There has been some research on such norm-guided behaviour in the NORMAS community of International Workshops on Normative Multi-Agent Systems.[5] See, for instance, [5].

4 The classic AI & law problems vs. the new challenge

For several reasons the above story about the practical applicability of AI & law research does not automatically apply to the problem of making autonomous systems obey the law. First, as we saw above, AI & law research has traditionally focused on support tools for humans carrying out legal tasks. With autonomous systems this is different: they do not support humans in their legal tasks (although they may support humans in other tasks) but they have to decide about the legal status of their own actions. In many cases it will be impossible for humans to check or override the system's decision.

Moreover, the tasks supported by traditional AI often concern the application of the law to past cases, to determine whether some past behaviour or some existing state of affairs is lawful or induces legal consequences. With autonomous systems this is different, since they have to think about the legal status of their future actions. Among other things, this means that in contrast to in traditional legal settings, autonomous systems do not face evidential problems in the legal sense. Even when traditional AI & law supports legal tasks with an eye to the future, such as deciding on benefit or permit applications, drafting regulations or contracts or designing tax constructions, there are differences with autonomous systems. While traditionally supported future-oriented task concern behaviour in the non-immediate future and often contain classes of actions (as with contract or with regulation design), autonomous systems have to 'run-time' consider individual actions in the immediate future.

Another difference, as explained in Section 3, is that the tasks supported by traditional AI & law are usually strictly legal while autonomous systems have to balance legal considerations against other considerations. This is not a black-and-white difference since, as explained in Section 2, law application also involves considering the social context and issues of fairness, common sense and the like. However, in the law, this is always done in service to the overall problem of classifying behaviour into legal categories. With autonomous systems this is different, since they do not have as their sole or primary aim to stay within the law.

Yet another difference is that the legal tasks supported by traditional AI & law tools require explanation and justification of decisions. With autonomous systems there is no need for this; all that counts is that legally acceptable behaviour is generated. Of course, when an autonomous system does something legally wrong, its behaviour might have to be explained in a court case. However, this does not require that the system itself can do that; it may suffice to have a log file recording the system's internal actions.

[4] http://www.bbc.com/news/technology-34808105, accessed 2 June 2016.

[5] http://icr.uni.lu/normas/, accessed 30 May 2016.

Finally, one may expect that the bulk of the cases encountered by an autonomous system will from a legal point of view be standard, mundane cases. For example, autonomous cars will not have to determine the legal responsibility for car accidents but will have to decide about driving from A to B in a way that respects the traffic regulations. While processing legislation in public administration also usually concerns standard cases, in the court room this is different.

5 Implications for applicability of AI & law research

What are the implications of the similarities and differences between the 'traditional' and new settings for the applicability of AI & Law research? The discussion here has to be speculative, since the answer depends on the type of autonomous system, how advanced it is, how safety-critical it is, and so on. Moreover, presently, there are still only few autonomous systems in practical use that have to take legally relevant decisions in a non-trivial way. Nevertheless, the technological developments go fast. Just 10 years ago, recent advances like IBM's Watson system and autonomously driving vehicles seemed unthinkable for the near future. Therefore, thinking about these issues cannot be postponed to the future.

Essentially, there have so far been three kinds of successful AI & law applications: decision support for large volumes of routine decision tasks (as in public administration); retrieval, summary and integration of legal information; and prediction of outcomes of decision problems in narrowly-defined factor-based domains.

Does the 'standard' nature of many cases faced by autonomous systems mean that the techniques for routine decision support as used in public administration can be applied to autonomous systems? This is not likely, since the traditional rule-based systems crucially rely on humans for preprocesses the input facts in legal terms and for overriding if necessary the system's decisions.

Can Watson-like legal research agents that retrieve, summarise and integrate information support autonomous systems? Here a similar problem arises, since the effective use of retrieved, summarised and integrated information still crucially relies on human judgement. Moreover, it remains to be seen whether the currently available legal information will be useful for the mundane and future-oriented normative decision problems faced by autonomous systems.

Are nonmonotonic reasoning techniques useful as a way to deal with exceptions and conflicting regulations? Not really, since such techniques do not offer ways to recognise the need for an exception to a legal rule or to recognize the best way to resolve a conflict between regulations, unless this has been programmed into the system in specific terms. Moreover, if the rules contain general exception clauses or the regulations contain general conflict resolution principles, the classification and interpretation problem will be too big.

Can machine-learning techniques as applied to factor-based domains support autonomous systems in classification and interpretation problems? Perhaps to some extent but there is room for caution here, since in the law these techniques have so far only worked for narrowly defined domains with a large amount of relatively consistent data. And the law does not have many of such domains. Moreover, when the data has to come from case law, a problem is that the cases may not be standard future-oriented cases of the kinds faced by the autonomous system. On the other hand, the 'traditional' drawback that these systems cannot justify or explain their output does not apply for autonomous systems, which are only meant to *generate* legally correct behaviour, not to explain or justify it.

Finally, there is the question whether an autonomous system should be designed to *reason* about how to behave lawfully or whether it can be *trained* to do so with machine-learning techniques applied to a large number of training cases. In the first approach there is the need for explicit representation of legal information in the system and for giving the system explicit reasoning and decision making capabilities. This is still somewhat similar to the traditional AI & law systems for supporting human decision making, except that the human is taken out of the loop. An important issue then is whether the mundane nature of cases faced by the autonomous system can reduce the complexity of the classification and interpretation problems to such an extent that the machine can fully take over. On the other hand, the reasoning can, unlike in the traditional settings, be opaque in that there is no need for explaining or justifying why the behaviour is legally correct. Incidentally, the latter combined with the run-time and forward-oriented setting with mundane cases, makes that the current research strands on evidential legal reasoning and sophisticated legal argument will likely be less relevant here.

The other approach is that the ability to behave legally correctly is acquired implicitly by training. For very advanced autonomous systems, like robots operating in daily life, this might be equivalent to solving the notorious AI common-sense problem, but for more modest systems this approach might do. One interesting question is how autonomous vehicles classify on this scale. [18] discuss some interpretation and classification problems in Dutch traffic law that are relatively easy for humans but seem very hard for the current generation of autonomous vehicles. The 'training' approach does not necessarily avoid the need for explicit representation of legal rules and regulations. They must now be represented as part of the design specification. One issue here is whether these specifications should be machine-processable in the same way as when designing explicit legal reasoners (as in the methods proposed by [3, 30]). It seems likely that at least some form of semi-formal representation is required, for purposes of verification and maintainability.

6 Conclusion

This position paper has been motivated by the rapidly increasing prospects of practically used autonomous artificial systems performing legally relevant tasks. The aim was to discuss how the current fruits of AI & law research on supporting human legal decision making can be used for making autonomous artificial systems behave lawfully. To this end the problems faced by human lawyers were compared to those faced by autonomous systems. The main similarity is that in both cases there is automated application of norms to facts. However, main differences are that the legal problems faced by autonomous systems have to be solved run-time and are future- instead of past-oriented. Moreover, while in traditional legal settings being lawful is the main goal, for autonomous systems it is only one of the concerns, to be balanced against, for example, social and individual goals. On the other hand, the legal problems faced by autonomous systems are, unlike those faced by lawyers in traditional settings, usually standard, mundane cases. Moreover, unlike lawyers in traditional settings, autonomous systems will usually not have to explain why their behaviour is lawful.

Because of the similarities, research on designing legally well-behaving autonomous systems can profit from the fruits of current AI & law research. However, because of the differences, applying these fruits in the new contexts is not trivial and requires extensive further research. In this position paper I have tried to create some awareness of the need for such research and pointed at some possible research directions.

REFERENCES

[1] K.D. Ashley, *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, Cambridge, MA, 1990.

[2] T.J.M. Bench-Capon, 'Neural networks and open texture', in *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, pp. 292–297, New York, (1993). ACM Press.

[3] T.J.M. Bench-Capon and F.P. Coenen, 'Isomorphism and legal knowledge based systems', *Artificial Intelligence and Law*, **1**, 65–86, (1992).

[4] T.J.M. Bench-Capon and H. Prakken, 'Argumentation', in *Information Technology and Lawyers: Advanced technology in the legal domain, from challenges to daily routine*, eds., A.R. Lodder and A. Oskamp, 61–80, Springer, Berlin, (2006).

[5] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre, 'Goal generation in the BOID architecture', *Cognitive Science Quarterly Journal*, **2**, 428–447, (2002).

[6] *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, eds., B.G. Buchanan and E.H. Shortliffe, Addison-Wesley, Reading, MA, 1984.

[7] A. Davis and G. Marcus, 'Commonsense reasoning and commonsense knowledge in artificial intelligence', *Communications of the ACM*, **58**, 92–103, (2015).

[8] J.H. de Wildt, *Rechters en Vage Normen. Een jurimetrisch onderzoek naar de uitleg van het begrip 'passende arbeid' in de Werkloosheidswet*, Gouda Quint, Arnhem, 1993. In Dutch. English title: Judges and vague norms. A jurimetric investigation of the interpetation of the concept 'suitable employment' in the Unemployment Act.

[9] R.M. Dworkin, 'Is law a system of rules?', in *The Philosophy of Law*, ed., R.M. Dworkin, 38–65, Oxford University Press, Oxford, (1977).

[10] T.F. Gordon, 'The Pleadings Game: an exercise in computational dialectics', *Artificial Intelligence and Law*, **2**, 239–292, (1994).

[11] T.F. Gordon and D.N. Walton, 'Proof burdens and standards', in *Argumentation in Artificial Intelligence*, eds., I. Rahwan and G.R. Simari, 239–258, Springer, Berlin, (2009).

[12] G. Governatori and G. Sartor, 'Burdens of proof in monological argumentation', in *Legal Knowledge and Information Systems. JURIX 2010: The Twenty-Third Annual Conference*, ed., R.G.F. Winkels, 37–46, IOS Press, Amsterdam etc., (2010).

[13] J.C. Hage, 'A theory of legal reasoning and a logic to match', *Artificial Intelligence and Law*, **4**, 199–273, (1996).

[14] P. Johnson. Legal knowledge-based systems in administrative practice and electronic service delivery. Tutorial notes, 13th Annual Conference on Legal Knowledge and Information Systems (JURIX 2000), 2000.

[15] P. Johnson and D. Mead, 'Legislative knowledge base systems for public administration - some practical issues', in *Proceedings of the Third International Conference on Artificial Intelligence and Law*, pp. 108–117, New York, (1991). ACM Press.

[16] A.J.I. Jones and M.J. Sergot, 'Deontic logic in the representation of law: towards a methodology', *Artificial Intelligence and Law*, **1**, 45–64, (1992).

[17] J. Lawrence and C. Reed, 'Combining argument mining techniques', in *Working Notes of the 2nd Argumentation Mining Workshop at ACL'2015*, Denver, CO, (2015).

[18] R.E. Leenes and F. Lucivero, 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design', *Law, Innovation and Technology*, **6**, 193–220, (2014).

[19] Ph. Morignot and F. Nashashibi, 'An ontology-based approach to relax traffic regulation for autonomous vehicle assistance', in *Proceedings of the 12th IASTED Conference on Artificial Intelligence and Applications*, Insbrück, Austria, (2013).

[20] E.W. Oskamp, *Computerondersteuning bij straftoemeting*, Gouda Quint, Arnhem, 1998. In dutch. English title: Computer Support of Sentencing.

[21] R. Mochales Palau and M.-F. Moens, 'Argumentation mining: the detection, classification and structure of arguments in text', in *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pp. 98–107, New York, (2009). ACM Press.

[22] H. Prakken, 'AI & law on legal argument: research trends and application prospects', *ScriptEd, A Journal of Law, Technology & Society*, **5**, 449–454, (2008). Editorial.

[23] H. Prakken, 'Legal reasoning: computational models', in *International Encyclopedia of the Social and Behavioural Sciences*, ed., J.D. Wright, Elsevier Ltd, Oxford, second edn., (2015).

[24] H. Prakken and G. Sartor, 'A dialectical model of assessing conflicting arguments in legal reasoning', *Artificial Intelligence and Law*, **4**, 331–368, (1996).

[25] H. Prakken and G. Sartor, 'A logical analysis of burdens of proof', in *Legal Evidence and Proof: Statistics, Stories, Logic*, eds., H. Kaptein, H. Prakken, and B. Verheij, 223–253, Ashgate Publishing, Farnham, (2009).

[26] H. Prakken and G. Sartor, 'Law and logic: A review from an argumentation perspective', *Artificial Intelligence*, **227**, 214–225, (2015).

[27] E.R. Rissland, K.D. Ashley, and R.P. Loui, 'AI and law: A fruitful synergy', *Artificial Intelligence*, **150**, 1–15, (2003).

[28] U. Schild, 'Criminal sentencing and intelligent decision support', *Artificial Intelligence and Law*, **6**, 151–202, (1998).

[29] J.S. Svensson, 'The use of legal expert systems in administrative decision making', in *Electronic Government: Design, Applications and Management*, ed., A. Grönlund, 151–169, Idea Group Publishing, London etc, (2002).

[30] T. van Engers, R. Gerrits, M. Boekenoogen, E. Glassée, and P. Kordelaar, 'POWER: Using UML/OCL for modeling legislation - an application report', in *Proceedings of the Eighth International Conference on Artificial Intelligence and Law*, pp. 157–167, New York, (2001). ACM Press.

[31] G.J. van Opdorp, R.F. Walker, J.A. Schrickx, C. Groendijk, and P.H. van den Berg, 'Networks at work. a connectionist approach to non-deductive legal reasoning', in *Proceedings of the Third International Conference on Artificial Intelligence and Law*, pp. 278–287, New York, (1991). ACM Press.

[32] B. Verheij, J.C. Hage, and H.J. van der Herik, 'An integrated view on rules and principles', *Artificial Intelligence and Law*, **6**, 3–26, (1998).

[33] M.G. Villalba and P. Saint-Dizier, 'Some facets of argument mining for opinion analysis', in *Computational Models of Argument. Proceedings of COMMA 2012*, eds., B. Verheij, S. Woltran, and S. Szeider, 23–34, IOS Press, Amsterdam etc, (2012).

Reified Input/Output logic - a position paper

Livio Robaldo and **Xin Sun**[1]

Abstract. We propose a new approach to formalize obligations and permissions from existing legislation. Specifically, we propose to combine two frameworks: Input/Output logic and the logic of prof. J.R. Hobbs. The former is a well-known framework in normative reasoning. The latter is a neo-Davidsonian wide-coverage first order logic for Natural Language Semantics. We propose to wrap Input/Ouput logic around Hobbs's logic, in order to fill the gap between current logical formalizations of legal text, mostly propositional, and the richness of Natural Language Semantics.

1 Introduction

State-of-the-art systems in legal informatics exploit NLP tools in order to transform, possibly semi-automatically, legal documents into XML standards such as Akoma Ntoso[2], where relevant information are tagged [5] [4]. Although these systems help navigate legislation and retrieve information, their overall usefulness is limited due to their focus on terminological issues while disregarding *semantic* aspects, which allow for legal reasoning.

Deontic Logic (DL) has been used since the 1950s as a formal instrument to model normative reasoning in law [38] [31]. However, subsequent developments in DL adopt an abstract view of law, with a very loose connection with the texts of regulations, which can be addressed with solutions coming from the literature on Natural Language Semantics (NLS). Most current proposals in DL are propositional, while NLS includes a wide range of fine-grained linguistic phenomena that require first-order logic (FOL) formalisms.

We aim at designing a logical framework able to fill the gap between standard (propositional) constructs used in DL and the richness of NLS. Among the logical frameworks (independently) proposed in the literature in NLS and DL respectively, we believe that two of them feature fundamental advantages: (1) the FOL of prof. J.R. Hobbs, designed to model the meaning of NL utterances via *reification*, and (2) Input/Output (I/O) logic, originally proposed in [27] to model deontic normative statements.

Reification is a concept originally introduced by the philosopher D. Davidson in [7]. It allows to move from standard notations in FOL such as '$(give\ a\ b\ c)$', asserting that 'a' gives 'b' to 'c', to another notation in FOL '$(give'\ e\ a\ b\ c)$', where e is the *reification* of the giving action. 'e' is a FOL term denoting the giving event by 'a' of 'b' to 'c'. In line with [2], e is said an "eventuality".

On the other hand, I/O logic is a well-known formalism in DL [9], thanks to its ability to deal with standard problems in DL, e.g., contrary-to-duty reasoning [27] and moral conflicts [32].

This paper presents a possible merging of Hobbs's logic and I/O logic that tries to combine their respective advantages. We restrict our attention to only *obligations* and *permissions*, i.e. the two main kinds of norms [36]. We leave other kinds of norms for future works.

We work on a corpus of EU directives, from which we selected the obligation in (1.a) (*Dir. 98/5/EC*) and the permission in (1.b) (*Dir. 2001/110/EC*). We did not find relevant differences between (1.a-b) and the other norms in the corpus, thus we assume our solution is general enough to cover a representative part of EU legislation.

(1) a. A lawyer who wishes to practise in a Member State other than that in which he obtained his professional qualification shall register with the competent authority in that State.

b. Where baker's honey has been used as an ingredient in a compound foodstuff, the term 'honey' may be used in the product name of the compound food instead of the term 'baker's honey'.

2 Related works

Some approaches in Legal Informatics try to model, in some deontic settings, NL sentences coming from *existing norms*, such as those in (1). The most representative work is perhaps [37]. Other examples may be found in [12] and [1]. Some approaches, e.g. [19], [8], and [15] among others, formalize legal knowledge via Event Calculus [23], a logical language extending reification by introducing special terms and predicates to deal with time points and time periods [10]. A similar account has been investigated by [22] in modal action logic.

To our knowledge, the approach that appears to be closest to the one we are going to propose below is perhaps McCarty's Language for Legal Discourse (LLD) [29]. LLD is strongly drawn on previous studies on NLS, it uses reification, and it aims at modeling existing legal text. [30] shows how it is possible to obtain LLD structures from federal civil cases in the appellate courts in USA via NLP.

However, LLD is very reminiscent of formalisms standardly used in NLS, such as Discourse Representation Theory (DRT) [21], and Minimal Recursion Semantics (MRS) [6]. Those are characterized by a close relation between syntax and semantics, in line with the well-known Montague's *principle of compositionality*[3], a cornerstone of standard formalisms used in NLS.

The principle of compositionality leads to representation based on *embeddings* of subformulae within the logical operators, which establish a *hierarchy* among the predications. For instance, a simple sentence like "John believes that Jack wants to eat an ice cream" could be represented via the following formula (assuming a de-dictio interpretation of the existential quantifier):

[1] University of Luxembourg, Luxembourg, {xin.sun, livio.robaldo}@uni.lu

[2] http://www.akomantoso.org

[3] http://plato.stanford.edu/entries/montague-semantics/#Com

$believe$[John,
$want$(Jack,
$\exists_x[(iceCream\ x) \wedge (eat$ Jack $x)]$]

Where $believe$ and $want$ are modal operators taking an individual as first argument and another (embedded) subformula as second argument. In the last formula, the operator $believe$ is hierarchically outscoping the operator $want$, in the sense that the latter occurs within the scope of the former.

Nevertheless, it has been shown by [16], [33], and [34] among others, that such an architecture prevents several available readings in NL, and more complex operators, able to connect the predications across the hierarchy, must be introduced to properly represent them.

For this reason, Hobbs proposed a logic where all formulae are *flat*, i.e. where no hierarchy is established among the predications.

3 Hobbs' logical framework

Prof. J.R. Hobbs defines a wide-coverage first-order logic (FOL) for NLS centered on reification. See [17] and several other earlier publications by the same author[4]. In Hobbs', eventualities may be *possible* or *actual*[5]. This distinction is represented via a predicate $Rexist$ that holds for eventualities really existing in the world. Eventualities may be inserted as parameters of such predicates as $want$, $believe$, etc. Reification can be applied recursively. The fact that "John believes that Jack wants to eat an ice cream" is represented as:

$$\exists_e\exists_{e1}\exists_{e2}\exists_x[\ (Rexist\ e) \wedge (believe'\ e\ \texttt{John}\ e_1) \wedge$$
$$(want'\ e_1\ \texttt{Jack}\ e_2) \wedge (eat'\ e_2\ \texttt{Jack}\ x) \wedge (iceCream'\ e_3\ x)\ ]$$

The crucial feature of Hobbs' logic, which distinguishes it from all other neo-Davidsonian approaches, e.g., LLD, is that all formulae are "flat", in the sense explained above. Specifically, the framework distinguishes between the formulae belonging to the ABox of an ontology from those belonging to its TBox. The ABox only includes conjunctions of atomic predicates asserted on FOL terms. On the other hand, the TBox defines these predicates in terms of the $Rexist$ predicate and standard FOL. All logical operators, e.g., boolean connectives[6], are modeled in this way. For instance, negation is modeled via a predicate not' defined in the TBox as:

(2) For all e and e_1 such that $(not'\ e\ e_1)$ holds, it also holds:
$$(Rexist\ e) \leftrightarrow \neg(Rexist\ e_1)$$

If $(not'\ e\ e_1)$ is true, all what we know is that the individuals e and e_1 are related via the not' predication. But this does not tell us anything about the real existence of either e or e_1. Similarly, and and $imply$ are "conjunctive" and "implicative" relations such that (3) and (4) respectively hold (on the other hand, we omit disjunction).

(3) For all e, e_1, e_2 such that $(and'\ e\ e_1\ e_2)$ holds, it also holds:
$$(Rexist\ e) \leftrightarrow (Rexist\ e_1) \wedge (Rexist\ e_2)$$

(4) For all e, e_1, e_2 such that $(imply'\ e\ e_1\ e_2)$ holds, it also holds:
$$(Rexist\ e) \leftrightarrow ((Rexist\ e_1) \rightarrow (Rexist\ e_2))$$

[4] See manuscripts at http://www.isi.edu/~hobbs/csk.html and http://www.isi.edu/~hobbs/csknowledge-references/csknowledge-references.html.

[5] Other approaches in the literature formalize this distinction in first-order logic, e.g. [3].

[6] See http://www.isi.edu/~hobbs/bgt-logic.text.

Hobbs and his followers implements a fairly large set of predicates for handling composite entities, causality, time, defeasibility, event structure, etc. For instance, [35] proposes a solution to model concessive relations, one of the most trickiest semantic relations occurring in NL, in Hobbs's logic.

The meaning of the predicates is restricted by adding 'axiom schemas'. Space constraints forbid us to illustrate details about all predicates defined by Hobbs. A possible axiom schema for the legal domain is shown in (5). (5) states that all lawyers are humans:

(5) For all e_1, x such that (*lawyer'* e_1 x) holds, it also holds:
$$\exists e_i\exists e_2[(imply'\ e_i\ e_1\ e_2) \wedge (Rexist\ e_i) \wedge (human'\ e_2\ x)]$$

4 Handling deontic defeasible reasoning in legal interpretation

A major problem in legal informatics concerns the proper interpretation of laws in given situations, which is up to the judges in courts [24]. *Legal interpretation* is a well-studied topic in legal informatics, cf. [25] among others. For instance, in (1.a), to what extent should we think of a lawyer who *wishes* to practise in a Member State different from the one he obtained his qualification? Under a *literal* interpretation of the verb "wishes", which may be taken as its default interpretation, a lawyer who simply *tells* some friends he would like to do so already violates the norm, if he is not registered with the competent authority. On the other hand, a reasonable (pragmatic) interpretation is that the norm is violated only if the non-registered lawyer performs some "formal" action, such as defending someone in court. According to the norm, that action should be blocked and the lawyer must possibly pay a penalty.

So far, few approaches have been proposed to handle multiple legal interpretations in logic. A recent one is [13], where a solution to deal with them in Defeasible Deontic Logic [12] via prioritized defeasible rules is proposed. Priorities are introduced to rank the available interpretations, i.e. to solve potential conflicts among them.

Following [13], we handle multiple legal interpretations via Hobbs's methodology to deal with defeasibility, which is in turn drawn from Circumscriptive Logic [28]. However, we do not claim that our solution features any particular advantage with respect to the one in [13], except the fact that our framework is first-order while Defeasible Deontic Logic is propositional.

The idea is simple and we illustrate it with an example. The fact that every bird flies is represented in FOL as $\forall_x[bird(x) \rightarrow fly(x)]$. In order to render the rule defeasible, we add another predicate $normalBF$ stating that birds fly only if it is "normal" to assume so: $\forall_x[(bird(x) \wedge normalBF(x)) \rightarrow fly(x)]$. Adding that emus are non-flying birds, i.e. $\forall_x[emu(x) \rightarrow (bird(x) \wedge \neg fly(x))]$, does not entail an inconsistency. It entails that $normalBF(x)$ is false for each emu x. In this sense, the latter rule is "stronger" than the former. Alternatively, we may directly assert that emus are not "normal" with respect to the property of flying, i.e. $\forall_x[emu(x) \rightarrow \neg normalBF(x)]$. $normalBF$ must be *assumed* to be true in order to trigger the property of flying on birds.

Different legal interpretations of "wishes" in (1.a) are similarly handled. Let us assume by default that if a lawyer x *says* he will practise in a Member State y, then he really wishes to do it.

(6) For all x, y, e_1, e_2, e_3 such that $(lawyer\ x) \wedge$ (*MS* y) $\wedge$ (*say'* e_1 x e_2) $\wedge$ (*wish'* e_2 x e_3) $\wedge$ (*practice'* e_3 x) $\wedge$ (*in* e_3 y) holds, it also holds:
$$\exists e_i[(imply'\ e_i\ e_1\ e_2) \wedge (Rexist\ e_i)]$$

To make (6) defeasible, we add a predicate *normalSP* stating that the entailment is valid only if it is "normal" to assume it:

(7) For all x, y, e_1, e_2, e_3 such that $(lawyer\ x) \wedge$ (*MS* y) $\wedge$ (*say'* $e_1\ x\ e_2$) $\wedge$ (*wish'* $e_2\ x\ e_3$) $\wedge$ (*practice'* $e_3\ x$) $\wedge$ (*in* $e_3\ y$) holds, it also holds:

$$\exists e_i \exists e_a \exists e_n [(imply'\ e_i\ e_a\ e_2) \wedge (Rexist\ e_i) \wedge (and'\ e_a\ e_1\ e_n) \wedge (\textit{normalSP'}\ e_n\ e_1)]$$

In (7), the real existence of e_1 is no longer sufficient to entail the one of e_2. In order to enable the entailment, the real existence of e_n is also needed. Now, a judge may reasonably decide that it is *not* normal assuming that a lawyer who says he will practice in a Member State entails that he "wishes" (in the sense of (1.a)) to do so, i.e.:

(8) For all x, y, e_1, e_2, e_3 such that $(lawyer\ x) \wedge$ (*MS* y) $\wedge$ (*say'* $e_1\ x\ e_2$) $\wedge$ (*wish'* $e_2\ x\ e_3$) $\wedge$ (*practice'* $e_3\ x$) $\wedge$ (*in* $e_3\ y$) holds, it also holds:

$$\exists e_n^n \exists e_n [(not'\ e_n^n\ e_n) \wedge (Rexist\ e_n^n) \wedge (\textit{normalSP'}\ e_n\ e_1)]$$

From (8), in case a lawyer x simply says he wishes to practice in a Member State y, we infer that e_n does *not* really exist. Thus, it is no longer possible to infer, from (7), whether e_2 really exists or not.

5 Input/Output logic

Input/Output (I/O) logic was introduced in [27]. It originates from the study of conditional norms. I/O logic is a family of logics, just like modal logic is a family of systems K, S4, S5, etc. However, unlike modal logic, which usually uses possible world semantics, I/O logic adopts *operational* semantics: an I/O system is conceived as a "deductive machine", like a black box which produces deontic statements as output, when we feed it factual statements as input.

As explained in [9], operational semantics solves the well-known Jørgensen's dilemma [20], which roughly says that a proper truth-conditional logic of norms is impossible because norms do not carry truth values. According to Jørgensen, typical problems of standard deontic logic arise from its truth-conditional model theory, i.e., possible world semantics. On the other hand, operational semantics straightforwardly allows to deal with contrary-to-duty reasoning, moral conflicts, etc. We address the reader to [26] and [32] among others for further explanations and examples.

Furthermore, I/O logic is one of the few existing frameworks for normative reasoning where also permissions, and not only obligations, have been studied in depth. Most current proposals are not specifically devoted to deal with existing legislation, and so they mostly focus on obligations only. For instance, in [15], devoted to handle business process compliance (BPC), obligations are analyzed in detail, while permissions are mostly neglected, in that the former play a role in BPC more prominent than the latter. The account in [15] has been recently extended to handle permissions in [11].

In [27], four basic I/O logics are defined: $\mathtt{out}_1$, $\mathtt{out}_2$, $\mathtt{out}_3$, and $\mathtt{out}_4$. Let L be standard propositional logic, let O and P be two subsets of $\mathtt{L} \times \mathtt{L}$, and let A to be a subset of L, i.e. a set of formulae in standard propositional logic. Each pair (a, b) in O is read as "given a, b is obligatory" while each pair (c, d) in P is read as "given c, d is permitted". Pairs in O and P are called "generators" and represent the "deduction machine": whenever one of the left-hand side (LHS) of the pairs is given in input, the corresponding right-hand side (RHS) is given in output.

(9) defines the semantics of $\mathtt{out}_1$, ... , $\mathtt{out}_4$. Cn is the consequence operator of propositional logic; it takes in input a set of formulae A and returns the set corresponding to the transitive closure of all formulae that can be entailed from A. A set of formulas is *complete* if it is either *maximally consistent* or equal to L.

(9)
- $\mathtt{out}_1(O,A) = Cn(O(Cn(A)))$
- $\mathtt{out}_2(O,A) = \bigcap\{Cn(O(V)) : A \subseteq V, V \text{ is complete}\}$
- $\mathtt{out}_3(O,A) = \bigcap\{Cn(O(B)) : A \subseteq B = Cn(B) \supseteq O(B)\}$
- $\mathtt{out}_4(O,A) = \bigcap\{Cn(O(V) : A \subseteq V \supseteq O(V)), V \text{ is complete}\}$

In (10), we report the axioms needed to define the I/O systems having the semantics from $\mathtt{out}_1$ to $\mathtt{out}_4$. $\vdash$ is the entailment relation of propositional logic.

(10)
- SI: from (a, x) to (b, x) whenever $b \vdash a$.
- OR: from (a, x) and (b, x) to $(a \vee b, x)$.
- WO: from (a, x) to (a, y) whenever $x \vdash y$.
- AND: from (a, x) and (a, y) to $(a, x \wedge y)$.
- CT: from (a, x) and $(a \wedge x, y)$ to (a, y).

The axioms in (10) constrain the generators belonging to O and P. For instance, CT says that in case two generators (a, x) and $(a \wedge x, y)$ belongs to O, then also the generator (a, y) *must* belong to O.

The derivation system based on SI, WO, and AND is called $\mathtt{deriv}_1$. Adding OR to $\mathtt{deriv}_1$ gives $\mathtt{deriv}_2$. Adding CT to $\mathtt{deriv}_1$ gives $\mathtt{deriv}_3$. The five rules together give $\mathtt{deriv}_4$. Each $\mathtt{deriv}_i$ is sound and complete with respect to $\mathtt{out}_i$ (see [27]).

An example of how the axioms in (10) work in practice is provided below directly on our FOL object logic. As pointed out above, the expressivity of I/O logic, as well as the one of its competitors, e.g., Imperative Logic [14], Prioritized Default Logic [18], and Defeasible Deontic Logic [12] among others, is limited to the *propositional* level. On the other hand, Hobbs's logic, thanks to its formal simplicity, allows to enhance the expressivity of I/O systems to the first-order level with little modifications of the axioms in (10).

6 Combining Input/Output logic and Hobbs's logic

Propositional logic does not have enough expressivity to represent real-world obligations and permissions, such as (1.a-b). Universally quantified variables and constant or functional terms are also needed.

For instance, "a lawyer" and "a Member State" in (1.a) refer to *every* lawyer and *every* Member State. On the other hand, the expression "that in which he obtained his professional qualification" ought to be represented as a function $f_1(x)$ that, given a lawyer x, returns the Member State where he obtained his professional qualification. Similarly, the expression "the competent authority in that State" is represented as a function $f_2(y)$ that, given a Member State y, returns the competent authority in that State. Finally, "the term 'honey' " and "the term 'baker's honey' " in (1.b) correspond to two FOL constants $\mathtt{T}_h$ $\mathtt{T}_{bh}$ respectively, denoting the two English words.

Our formulae are Hobbs's conjunctions of atomic predications, possibly involving FOL variables. Some of those variables will occur both in the LHS and the RHS of an I/O generator, while the others will occur either in the LHS or in the RHS. The variables occurring in both will be universally quantified, while the ones occurring in either one of the two will be existentially quantified. Furthermore, we will require each formula of the object logic to assert exactly one $Rexist$ predicate on the main eventuality. As explained in section 3, the semantics of Hobbs's logic is centered on the $Rexist$ predicate.

We add a single construct to the syntax of the generators: universal quantifiers for binding the variables occurring in both the LHS and the RHS. These quantifiers act as "bridges" between the LHS and the RHS, in order to "carry" individuals from the input to the output. Formally, our generators have the following form, where $LHS(x_1, x_2, \ldots, x_n)$ and $RHS(x_1, x_2, \ldots, x_n)$ are conjunctions of FOL predicates; $x_1, x_2, \ldots, x_n$ are free in LHS and RHS but they are externally bound by universal quantifiers. LHS and RHS will possibly include other existentially quantified variables.

$$\forall_{x_1} \forall_{x_2} \ldots \forall_{x_n} (LHS(x_1, x_2, \ldots, x_n),\ RHS(x_1, x_2, \ldots, x_n))$$

This architectural choice is motivated by an empirical analysis of the obligations/permissions in our corpus of EU Directives. Norms found in legislation typically hold for all members in a certain set of individuals, e.g. the set of all lawyers. On the other hand, we did not find in our corpus any obligation or permission in the form "If a lawyer exists, then he is obliged to take some actions". This sounds quite intuitive: statements in legislation are typically *universal* assertions, i.e., they do not hold for single specific individuals.

Note that, in any case, as long as formulae are conjunctions of atomic predicates, de re obligations/permissions can be easily dealt with by removing existentials via skolemization. A generator in the form $\exists_x$(LHS(x), $RHS(x)$) can be substituted by (LHS(i), $RHS(i)$), where i is a FOL constant skolemizing $\exists_x$. On the other hand, a generator in the form $\forall_x \exists_y$(LHS(x, y), $RHS(x, y)$) can be substituted by $\forall_x$(LHS(x, $f(x)$), $RHS(x, f(x))$), where f is a FOL function skolemizing $\exists_x$. Existentials occurring in the object logic formulae can be also skolemized. For instance, a generator in the form $\forall_x(\exists_y LHS(x, y)$, $RHS(x))$ can be substituted by $\forall_x$(LHS(x, $f(x)$), $RHS(x)$), where f is a FOL function skolemizing $\exists_x$.

Similarly, it must be observed that, in finite domains, universal quantifiers are just a compact way to refer to all individuals in the universe. We obtain an equivalent set of generators by substituting the universally quantified variables with all constants referring each to an individual in the universe. For instance, assuming the universe includes the individuals a, b, c only, the generator $\forall_x$(LHS(x), $RHS(x)$) is equivalent to the set of generators (LHS(a), $RHS(a)$), (LHS(b), $RHS(b)$), and (LHS(c), $RHS(c)$).

6.1 Generalizing Input/Output logic axioms

We have proposed above to integrate Hobbs's logic within I/O generators by simply adding wide-scope universal quantifiers to the syntax of the generators, in order to create a "bridge" for "carrying" the FOL terms matching the LHS to the output. Also the axioms in (10) need to be generalized accordingly. This section shows the generalization of the axiom CT. The generalization of the other axioms is similar and it is left to the reader as an exercise. CT is generalized as in (11).

(11) from: $\forall_{x_1} \ldots \forall_{x_n}($
$\exists_{e_{11}} \exists_{y_{11}} \ldots \exists_{y_{1i}} [(Rexist\ e_{11}) \wedge (\Psi'_1\ e_{11}\ y_{11} \ldots y_{1i}\ x_1 \ldots x_n)],$
$\exists_{e_{21}} \exists_{y_{21}} \ldots \exists_{y_{2j}} [(Rexist\ e_{21}) \wedge (\Psi'_2\ e_{21}\ y_{21} \ldots y_{2j}\ x_1 \ldots x_n)])$
and: $\forall_{x_1} \ldots \forall_{x_n} (\exists_e \exists_{e_{11}} \exists_{y_{11}} \ldots \exists_{y_{1i}} \exists_{e_{21}} \exists_{y_{21}} \ldots \exists_{y_{2j}} [$
$(Rexist\ e) \wedge (and'\ e\ e_{11}\ e_{21}) \wedge (\Psi'_1\ e_{11}\ y_{11} \ldots y_{1i}\ x_1 \ldots x_n) \wedge$
$(\Psi'_2\ e_{21}\ y_{21} \ldots y_{2j}\ x_1 \ldots x_n)],$
$\exists_{e_{31}} \exists_{y_{31}} \ldots \exists_{y_{3k}} [(Rexist\ e_{31}) \wedge (\Psi'_3\ e_{31}\ y_{31} \ldots y_{3k}\ x_1 \ldots x_n)])$
to: $\forall_{x_1} \ldots \forall_{x_n}($
$\exists_{e_{11}} \exists_{y_{11}} \ldots \exists_{y_{1i}} [(Rexist\ e_{11}) \wedge (\Psi'_1\ e_{11}\ y_{11} \ldots y_{1i}\ x_1 \ldots x_n)],$
$\exists_{e_{31}} \exists_{y_{31}} \ldots \exists_{y_{3k}} [(Rexist\ e_{31}) \wedge (\Psi'_3\ e_{31}\ y_{31} \ldots y_{3k}\ x_1 \ldots x_n)])$

An example is: given "Every lawyer is obliged to run" and "Every lawyer who runs is obliged to wear a red hat", formalized in (12):

(12) $\forall_x(\ \exists_{e_{11}} [(Rexist\ e_{11}) \wedge (lawyer'\ e_{11}\ x)],$
$\exists_{e_{21}} [(Rexist\ e_{21}) \wedge (run'\ e_{21}\ x)]\ )$

$\forall_x(\ \exists_e \exists_{e_{11}} \exists_{e_{21}} [(Rexist\ e) \wedge (and'\ e\ e_{11}\ e_{21}) \wedge$
$(lawyer'\ e_{11}\ x) \wedge (run'\ e_{21}\ x)],$
$\exists_{e_{31}} [(Rexist\ e_{31}) \wedge (wearRedHat'\ e_{31}\ x)]\ )$

in case the I/O system includes the axiom in (11), O must include (13), which refers to "Every lawyer is obliged to wear a red hat".

(13) $\forall_x(\ \exists_{e_{11}} [(Rexist\ e_{11}) \wedge (lawyer'\ e_{11}\ x)],$
$\exists_{e_{31}} [(Rexist\ e_{31}) \wedge (wearRedHat'\ e_{21}\ x)])$

6.2 Formalizing the examples in (1)

We have now all the ingredients for representing (1.a-b). In a normative Input/Output system N=(O,P), the former is inserted in O while the latter is inserted in P. The formula representing (1.a) is:

(14) $\forall_x \forall_y(\ \exists_{e_1} \exists_{e_2} [(Rexist\ e_1) \wedge (lawyer\ x) \wedge$ (*MS* y) $\wedge$
(*wish'* $e_1\ x\ e_2$) $\wedge$ (*practice'* $e_2\ x$) $\wedge$ (*in* $e_2\ y$) $\wedge$ *diffFrom*($y\ f_1(x)$)],
$\exists_{e_3} [(Rexist\ e_3) \wedge$ (*register'* $e_3\ x$) $\wedge$ (*at* $e_3\ f_2(y)$)])

As discussed in section 4, the predicate *wish*, as well as any other predicate, may be subject to different legal interpretations, which may be asserted in the knowledge base via the mechanism used in Hobbs's to handle defeasibility.

The permission in (1.b) is similarly formalized as in (15).

(15) $\forall_y(\exists_x \exists_{e_1} [(Rexist\ e_1) \wedge$ (*ingrOf'* $e_1\ x\ y$) $\wedge$
(*bakerHoney* x) $\wedge$ (*foodStuff* y)],
$\exists_{e_2} [(Rexist\ e_2) \wedge$ (*substitute*' $e_2\ \mathtt{T}_h\ \mathtt{T}_{bh}$) $\wedge$ (*in* $e_2\ f_3(y)$)]

Note that the variable x occurs in the LHS only, thus it is existentially quantified. The formula in (15) reads as follows: for each compound foodstuff y for which it is "true" (in the sense that it really exists in the current world) the fact that one of its ingredients is baker's honey, then it is permitted that, in the current world, also the fact that the term 'honey' is substituted by the term 'baker's honey' in the product name of y really exist.

ACKNOWLEDGEMENTS

Livio Robaldo has received funding from the European Unions H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 661007 for the project "ProLeMAS: PROcessing LEgal language in normative Multi-Agent Systems". Xin Sun has received funding from the European Union's H2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

REFERENCES

[1] *Logic in the theory and practice of lawmaking*, eds., M. Araszkiewicz and K. Pleszka, Springer, 2015.

[2] E. Bach, 'On time, tense, and aspect: An essay in english metaphysics', in *Radical Pragmatics*, ed., P. Cole, Academic Press, New York, (1981).

[3] P. Blackburn and J. van Benthem, 'Modal logic: A semantic perspective', in *Handbook of Modal Logic*, eds., P. Blackburn, J. van Benthem, and F. Wolter, Elsevier, (2007).

[4] G. Boella, L. Di Caro, and L. Robaldo, 'Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines', in *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pp. 218–225. Springer, (2013).

[5] G. Boella, L. Di Caro, A. Ruggeri, and L. Robaldo, 'Learning from syntax generalizations for automatic semantic annotation', *Journal of Intelligent Information Systems*, **43**(2), 231–246, (2014).

[6] A. Copestake, D. Flickinger, and I.A. Sag, 'Minimal Recursion Semantics. An introduction', *Journal of Research on Language and Computation.*, **2**(3), (2005).

[7] D. Davidson, 'The logical form of action sentences', in *The Logic of Decision and Action*, ed., N. Rescher, Univ. of Pittsburgh Press, (1967).

[8] A. Farrell, M. Sergot, M. Salle, and C. Bartolini, 'Using the event calculus for tracking the normative state of contracts', *International Journal of Cooperative Information Systems*, **14**, 99–129, (2005).

[9] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. (eds.) van der Torre, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013.

[10] Antony Galton, 'Operators vs. arguments: The ins and outs of reification', *Synthese*, **150**(3), 415–441, (2006).

[11] G. Governatori and M. Hashmi, 'Permissions in deontic event-calculus', in *International Conference on Legal Knowledge and Information Systems (Jurix)*, pp. 181–182, Braga, Portugal, (2015).

[12] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco, 'Computing strong and weak permissions in defeasible logic', *Journal of Philosophical Logic*, **6**(42), (2013).

[13] G. Governatori, A. Rotolo, and G. Sartor, 'Deontic defeasible reasoning in legal interpretation', in *The 15th International Conference on Artificial Intelligence & Law*, ed., K. Atkinson, San Diego, (2015).

[14] Jörg Hansen, 'Prioritized conditional imperatives: problems and a new proposal', *Autonomous Agents and Multi-Agent Systems*, **17**(1), (2008).

[15] M. Hashmi, G. Governatori, and M. Wynn, 'Modeling obligations with event-calculus', in *Rules on the Web. From Theory to Applications*, eds., A. Bikakis, P. Fodor, and D. Roman, volume 8620 of *Lecture Notes in Computer Science*, 296–310, Springer International Publishing, (2014).

[16] J. R. Hobbs, 'Deep lexical semantics', in *Proc. of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, Haifa, Israel, (2008).

[17] J.R. Hobbs, 'The logical notation: Ontological promiscuity', in *Chap. 2 of Discourse and Inference*, (1998). Available at http://www.isi.edu/∼hobbs/disinf-tc.html.

[18] John Horty, *Reasons as Defaults*, Oxford University Press, 2012.

[19] Andrew J.I. Jones and Steven Orla Kimbrough, *A Note on Modelling Speech Acts as Signalling Conventions*, 325–342, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[20] Jorgen Jørgensen, 'Imperatives and logic', *Erkenntnis*, **7**, (1937).

[21] H. Kamp and U. Reyle, *From Discourse to Logic: an introduction to model-theoretic semantics, formal logic and Discourse Representation Theory*, Kluwer Academic Publishers, Dordrecht, 1993.

[22] Samit Khosla and T. S. E. Maibaum, 'The prescription and description of state based systems', in *Temporal Logic in Specification, Altrincham, UK, April 8-10, 1987, Proceedings*, eds., Behnam Banieqbal, Howard Barringer, and Amir Pnueli, volume 398 of *Lecture Notes in Computer Science*, pp. 243–294. Springer, (1987).

[23] R Kowalski and M Sergot, 'A logic-based calculus of events', *New Generation Computing*, **4**(1), 67–95, (1986).

[24] Doris Liebwald, 'Vagueness in law: A stimulus for 'artificial intelligence & law", in *Proc. of the 14th International Conference on Artificial Intelligence and Law*, ICAIL '13, pp. 207–211, (2013).

[25] N. MacCormick and R.S. Summers, *Interpreting Statutes: A Comparative Study*, Applied legal philosophy, Dartmouth, 1991.

[26] David Makinson and Leendert van der Torre, 'Constraints for input/output logics', *Journal of Philosophical Logic*, **30**(2), (2001).

[27] David Makinson and Leendert W. N. van der Torre, 'Input/output logics', *Journal of Philosophical Logic*, **29**(4), 383–408, (2000).

[28] J. McCarthy, 'Circumscription: A form of nonmonotonic reasoning', *Artificial Intelligence*, (13), 27–39, (1980).

[29] L. T. McCarty, 'Ownership: A case study in the representation of legal concepts', *Artificial Intelligence and Law*, **10**(1-3), 135–161, (2002).

[30] L. T. McCarty, 'Deep semantic interpretations of legal texts', in *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pp. 217–224, (2007).

[31] J.J. Meyer and R.J. Wieringa, *Deontic Logic in Computer Science: Normative system specification.*, John Wiley and sons Ltd, 1993.

[32] X. Parent, 'Moral particularism in the light of deontic logic', *Artificial Intelligence and Law*, **19**(2-3), 75–98, (2011).

[33] L. Robaldo, 'Interpretation and inference with maximal referential terms.', *The Journal of Computer and System Sciences*, **76(5)**, (2010).

[34] L. Robaldo, 'Conservativity: a necessary property for the maximization of witness sets.', *The Logic Journal of the IGPL*, **21(5)**, 853–878, (2013).

[35] L. Robaldo and E. Miltsakaki, 'Corpus-driven semantics of concession: Where do expectations come from?', *Dialogue&Discourse*, **5**(1), (2014).

[36] G. Sartor and E. Pattaro, *Legal Reasoning: A Cognitive Approach to the Law*, Treatise of legal philosophy and general jurisprudence, 2005.

[37] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory, 'The british nationality act as a logic program', *Communications of the ACM*, **29**(5), (1986).

[38] G. H. von Wright, *An Essay in Deontic Logic and the General Theory of Action*, Acta Philosophica Fennica, Fasc. 21, North-Holland, 1968.

Recognizing Cited Facts and Principles in Legal Judgements

Olga Shulayeva and **Advaith Siddharthan** and **Adam Wyner** [1]

Abstract. In common law jurisdictions, legal professionals cite facts and legal principles from precedent cases to support their arguments before the court for their intended outcome in a current case. This practice stems from the doctrine of *stare decisis*, where cases that have similar facts should receive similar decisions with respect to the principles. It is essential for legal professionals to identify such facts and principles in precedent cases, though this is a highly time intensive task. In this paper, we present studies that demonstrate that human annotators can achieve reasonable agreement on which sentences in legal judgements contain cited facts and principles (respectively, $\kappa = 0.65$ and $\kappa = 0.95$ for inter- and intra-annotator agreement). We further demonstrate that it is feasible to automatically annotate sentences containing such legal facts and principles in a supervised machine learning framework, reporting per category precision and recall figures of between 79% and 89% for classifying sentences in legal judgements as cited facts, principles or neither using a Bayesian classifier, with an overall κ of 0.72 with the human-annotated gold standard.

1 Introduction

In common law jurisdictions, legal practitioners treat existing case decisions (precedents) as a source of law. Case citations, references to legal precedents, are an important argumentation tool, enabling lawyers to formulate and present their argument persuasively. This practice stems from the doctrine of *stare decisis*, which can be translated from Latin as to 'stand by the decided cases' [2], where a case under consideration that has facts similar enough to precedent cases should receive similar decisions as the precedents. A legal professional looks to establish the relevant law in the current case; to do so, she must consult precedent cases in order to establish how similar patterns of facts were decided. Citations from existing case law are used to illustrate legal principles and facts that define the conditions for application of legal principles in the current case.

Citation analysis can help legal practitioners to identify which principles have applied in a certain case and which facts have been selected as the 'material' facts of the case, i.e. the facts that influenced the decision and which are crucial in establishing the similarity between two cases. There is no defined guide on how to identify the law embedded within common law decisions, so legal professionals are expected to make themselves familiar with as many relevant decisions as possible in order to make informed predictions about the outcome of a current case. Decisions delivered by courts are binding and can therefore provide useful information for legal professionals. The information that is embedded within the cited cases includes the legal principles and facts that are used to reason to a decision. Optimally, a legal professional finds a cited case with the same facts and legal principles, and so can argue that the decision for the current case should be that of the precedent; similarly, the opposing party may identify precedents with opposing principles to argue the decision should be otherwise. More commonly, legal professionals must consider a range of precedents, each of which highlight particular facts and legal principles that support their argument (or argue against the opposition). It is, then, essential that each side in the legal dispute identifies a relevant case base which supports the legal claims made during legal arguments. As the body of common law is continually growing, human citation analysis is complex as well as knowledge and time intensive.

To support citation analysis (discussed further in Section 2.1), existing electronic tools, such as electronic databases[3], provide one word summaries for relationships between cases (e.g.'applied'). However, it is uncommon for them to extract information about the facts and the legal principles of the cited cases. This means that on many occasions readers are required to make themselves familiar with the full text of multiple law reports in order to identify the applicable law and the correct way to apply it. Thus, citation analysis tools save some labour by providing a preliminary filter on relevant cases, yet, identification of particular cases and the essential details require further manual effort.

In the course of working on citation analysis, certain key concepts of legal theory must be scoped, given that this is a report on the computational analysis of the language of the law rather than on legal theory. In particular, cases are considered to contain *ratio decidendi*, which can be translated as a *reason for a decision*, an important piece of reasoning that is incorporated into the argumentation structure of future decisions. A variety of approaches to defining *ratio decidendi* can be identified in legal theory. As defined by [28]: '*ratio decidendi* can be identified as those statements of law which are based on the facts as found and upon which the decision is based'. [11] provides several explanations on what forms the binding part of a decision:

> '(1) the rule(s) of law that the court explicitly states, or that can reasonably be inferred, that it regarded as necessary to (or important in) its resolution of the case [. . .], (2) facts the precedent court regarded as 'material,' i.e., crucial for the court's resolution, plus the result of the case; and (3) facts the court now constrained by the precedent regards as material in the earlier case plus its result.'

The complexities stemming from the debates surrounding the defi-

[1] Department of Computing Science, University of Aberdeen, United Kingdom, email: {olga.shulayeva,advaith,azwyner}@abdn.ac.uk

[2] Source: http://thelawdictionary.org/

[3] e.g. LexisNexis Shepard's Citations Service `http://www.lexisnexis.com/en-us/products/shepards.page`

nition of ratio are excluded from the scope of this paper. Here, *ratio* will be understood as a combination of the facts of the current case along with the legal principles that are invoked when the facts of the current case are similar enough to the facts of the case that established the precedent.

This paper makes a novel, preliminary contribution towards automated identification of legal principles and facts embedded within common law citations. A gold standard corpus is created, with sentences containing legal principles and facts manually annotated. A Bayesian Multinomial Classifier (using Weka) is then applied to the corpus using a set of linguistic features to automatically identify these sentences. The main results are a demonstration that (a) the human annotation task is feasible, i.e. human annotators can achieve reasonable agreement on which sentences in legal judgements contain cited facts and principles and (b) it is feasible to automatically annotate sentences containing such legal facts and principles to a high standard. The reported studies lay the basis for further applications, including creation of meta-data for search and retrieval purposes, compilation of automated case treatment tables containing summaries about legal principles and material facts of cases, and automated analysis of reasoning patterns and consistency applied in legal argumentation.

We first present related work in Section 2. Then there are two studies, on manual annotation in Section 3 and on automated annotation in Section 4. The paper closes with some conclusions in Section 5.

2 Related work

This research aims to apply machine learning methodology in order to automatically identify legal principles and facts in case citations. A significant amount of work has been done in the area of citation analysis in scientific literature, while only a very small amount of work has been done that focuses on studying case law citations. Most existing studies on case law citations aim to identify case treatment – the relationship between citing and cited cases (e.g. *distinguished*, *explained*, and others) – or analyse citations from the point of view of network analysis, but don't focus on fine-grained analysis of the cited information. To the best of our knowledge, there is no reported work that specifically aims to apply machine learning methodology to identify legal principles and facts of the cited cases in case citations. In the following subsections, we discuss related work on citation analysis along with relevant literature on legal argumentation.

2.1 Citation analysis

The first attempts to systematise citation information were done in the field of common law by the developers of legal citators, starting with Frank Shepard in 1873, who relied on human expertise to provide discourse-aware summaries of case law citations. More recently, citation information is presented as in LexisNexis Shepard's Citations Service.

Despite lawyers being the pioneers of citation analysis [26], the research on citation analysis in common law has not been developing as fast as citation analysis in the domain of scientific reports. Eugene Garfield is often cited as one of the pioneers and key contributors towards citation analysis in science. Garfield was inspired by the Shepard's citations and argued that similar methodologies can be useful for summarisation of scientific citations [10]. Garfield employed a bibliographic approach to create ICI Citation Indexes, and the data from citation indexes was later used for a number of bibliometric studies that "extract, aggregate and analyse quantitative aspects of bibliographic information" [22]. He believed that citation analysis could be used for evaluation of scientific performance, for example, in calculation of journal ranks based on citation frequency and impact. As noted by [22], quantitative data from bibliometric studies is widely used to assess the performance of individual scholars, scientific journals, research institutions and 'general, structural aspects of the scholarly system' (e.g. measuring trends in national publication output). [22] also concluded that ICI citation indexes do not 'capture motives of individuals, but their consequences at an aggregate level' and argued for further development of qualitative citation based indicators, thus abandoning the principle underlying most citation analyses that 'all citations are equal'. Qualitative approaches in citation analysis take into account the intentions of the person who was providing the citation. They aim to capture citation qualities that are overlooked by quantitative methodologies, for example, such as polarity and sentiment. A scientific article may be frequently cited, but it can be due to criticisms or mere acknowledgements, which distinguishes it from an article introducing an approach that is widely accepted and utilised. Several researchers can be mentioned in respect of qualitative citation based indicators in science [24, 29, 4, 32, 2]. [6] conducted a research of citation behaviours and noted that at the time there was not a universal approach in citation studies. Application of qualitative citation based indicators often relies on linguistic discourse markers to generate conclusions about citations and citing behaviours. For example, citations can be classified according to sentiment polarities: confirmative or negative [24]; positive, neutral or weak [32].

Recently there has been more interest toward citation studies in law, where there appear to be two major directions: applying network analysis to citations [37, 19, 34, 20, 33, 25] and classification systems allowing one to estimate the 'treatment' status of the cited case [16, 9].

[37] developed Semantics-Based Legal Citation Network, a tool that extracts and summarises citation information, allowing the users to 'easily navigate in the citation networks and study how citations are interrelated and how legal issues have evolved in the past.' The researchers note that different parts of a case can be cited and studying the reasons for citation can provide valuable information for a legal researcher. Their approach relied on RFC (reason for citing), a patented technology that allows extracting reasons of why the case has been cited. RFC performance was summarised in the patent [15], and it explored a methodology of 'identifying sentences near a document citation (such as a court case citation) that suggest the reason(s) for citing (RFC)'. In [37], the information retrieved by RFC was further organised into semantic citation networks. The task of identifying RFC may be somewhat similar to the task that is undertaken as a part of this project due to the fact that information contained in principles and facts of cited cases can be used as a part of estimating reasons for citing.

History Assistant was designed by [16] to automatically infer direct and indirect treatment history from case reports. Direct treatment history covered historically related cases, such as appeals etc. Indirect treatment history dealt with the cited cases within a document in order to establish how the cited case has been treated. It relied on the classification methodology of Shepard's citations that combines the knowledge about sentiment and aims of legal communication with heuristic information about court hierarchy. It includes such classes as applied, overruled and distinguished. History Assistant was expected to be an aid for editorial work rather than replace the effort of the editors. The program consisted of a set of natural language modules and a prior case retrieval module. Natural language process-

ing relied on machine learning methodology and employed statistical methods over annotated corpus.

[9] created LEXA – a system that relied on RDR (Ripple Down Rules) approach to identify citations within the 'distinguished' class. This category is generally best linguistically signalled and is therefore suitable for achieving high precision and recall. The key idea underpinning RDR was that the 'domain expert monitors the system and whenever it performs incorrectly he signals the error and provides as a correction a rule based on the case which generated the error, which is added to the knowledge base' [9]. The approach employed annotators to create an initial set of rules leaving the end users to refine and further expand the set. The authors claimed that 'the user can at any stage create new annotations and use them in creating rules' which may put a more significant reliance on the user input than an end user may be equipped or expecting to provide. LEXA employed 78 rules that recognized 'distinguished' citations with a precision of 70% and recall of 48.6% on the cleaned test set, which is significantly lower than the results reported by [16] for the same category: precision (94%) and recall (90%). The difference in results suggests that a complex fine-grained analysis used by [16] that included machine-learning for language processing may help achieve better classification outcomes.

2.2 Argument extraction

There have been a variety of attempts aimed at automated extraction of argumentation structure of text and its constituents. The methodologies employed by such studies often rely on extraction and further analysis of linguistic information that is available within the text. One of the relatively recent successful examples of argumentation extraction methodology can be argumentation zoning. This approach is based on the assumption that the argumentation structure can be presented as a combination of rhetorical zones that are used to group the statements according to their rhetorical role. This approach was initially used by [30, 31] for scientific reports. [12] used argumentation zoning to create summaries for common law reports. Both studies report acceptable results for most of the categories, with some categories performing better than others.

An approach similar to argumentation zoning was taken by [8] to develop a scheme for identification of argument structure of Canadian case law and [18] to analyse the structure of German court decisions. A methodology relying on manual annotation of discourse structures and in that respect similar to argumentation zoning was used by [36] to detect case elements such as Case citation, cases cited, precedential relationships, Names of parties, judges, attorneys, court sort, Roles of parties (i.e. plaintiff or defendant), attorneys, and final decision. Whilst the methodology developed does not aim to fully reconstruct argumentation structure, the information obtained during the study can be used as a part of a wider application.

[35] conducted a study aimed at identification of argumentation parts with the use of context-free grammars. Similar to [16] the study reports the following difficulties with identifying argumentation structures in legal texts: '(a) the detection of intermediate conclusions, especially the ones without rhetorical markers, as more than 20% of the conclusions are classified as premises of a higher layer conclusion; (b) the ambiguity between argument structures.' The results reported are as follows: premises – 59% precision, 70% recall; conclusions – 61% precision, 75% recall; non-argumentative information – 89% precision, 80% recall.

The methodology of applying statistical tools over annotated corpus was employed by [23] to automatically detect sentences that are a part of the legal argument. The study achieved 68% accuracy for legal texts. [1] aimed to extract 'argumentation-relevant information automatically from a corpus of legal decision documents' and 'build new arguments using that information'.

A related, important distinction that should be made with regard to legal argumentation is the idea that the cited legal principles can be classed as ratio or obiter. As defined by [28]: 'ratio decidendi can be understood as those statements of law which are based on the facts as found and upon which the decision is based'. Statements that are usually included into obiter class are dissenting statements and statements that are 'based upon either nonexistent or immaterial facts of the case' [28]. From the point of view of law the main difference between ratio and obiter is that the former is binding, while the latter only possesses persuasive powers. [3] tried to automatically identify and extract ratio. [27] tried to identify obiter statements. However, the distinctions between ratio or obiter will not be used as a part of this work.

3 Manual annotation study

The manual annotation study focused on annotating the gold standard corpus and evaluating the annotation methodology. This gold standard corpus was used to extract the features necessary for the machine annotation study. Two annotators were used for the purposes of the manual annotation study: Annotator 1 and Annotator 2. Annotator 1 has legal training and Annotator 2 does not. All manual annotation was performed in GATE[4].

3.1 Method

The corpus for the gold standard was compiled from 50 common law reports that had been taken from the British and Irish Legal Institute (BAILII) website in RTF format. The length and structure of reports varied, which was most often defined by the complexity of the matter: longer and more complicated cases often had more sections. As reported by GATE Sentence Splitter (GATE 8.0.), the full corpus contained 1211012 tokens (or words) and 22617 sentences which included headings and other units that didn't form full sentences from grammatical point of view. Most reports had a section on the top introducing the court, the parties, legal representatives, case number etc. It was often the case that the legal situation was presented in the introduction and that the legal analysis was in the middle of the report. However, the reports did not follow a universal format. Conclusions were often short and situated at the end of the report. Case law citations are used to support legal argumentation and are therefore referred to as a part of legal analysis. For that reason they were rarely found in introduction or conclusion.

Annotator 1 created annotation guidelines (high level task definition, descriptions and examples for each category, and analyses of a few difficult cases) in several iterations and trained Annotator 2 on their use. The annotators were expected to identify sentences that contained the legal *principles* and *facts* of the cited cases, based on the written guidelines. Sentences associated with cited cases that are neither *principles* or *facts* are annotated as *neutral*.

The task of annotation focused on the identification of cited information within annotation areas that were defined as paragraphs having at least one citation. Citation instances had been manually annotated prior to the study. Given the discussion of the complexity

[4] GATE 8.0: https://gate.ac.uk

of jurisprudential views of legal principles, we have taken an operationalised view, based on the analysis of a legal scholar and key linguistic indicators.

All propositions that are associated with the cited case should annotated if the court deems they support the legal reasoning of the citing case. A *legal principle* is a statement which is used, along with facts, to reach a conclusion. Linguistically, a legal principle can for instance be indicated by deontic modality, e.g. expressions of *must* for obligation, *must not* for prohibition, or *may* for permission, which contrast with epistemic modalities for necessity and possibility. For example:

> As a matter of principle no order should be made in civil or family proceedings without notice to the other side unless there is a very good reason for departing from the general rule that notice must be given. (Gorbunova v Berezovsky (aka Platon Elenin) & Ors, 2013)

Legal principles can be qualified, e.g. with conditions that may limit the application of rule. It is also possible that legal principles are "active" in reasoning, yet inferred from the text, in which case, they cannot be annotated or used for further text processing.

In contrast to legal principles, there are *facts*, which are statements bearing on what uncontroversially exists, occurred, or is a piece of information. For our purposes, only sentences that refer to events which occur outside the court hearing are annotated; this excludes procedural facts. For example:

> Miss Lange was not a party to the 1965 Transfer or the 1968 Deed and she covenanted only with Mrs de Froberville (and not with Brigadier Radford) to comply with the covenants in those instruments in so far as they were still subsisting and capable of taking effect. (89 Holland Park (Management) Ltd & Ors v Hicks, 2013)

Linguistically, facts present themselves with non-modal expressions and denoting expressions, e.g. are not generic, non-actual, and indefinite.

Following a period of training, a set of 10 reports were randomly selected (all previously unseen by the annotators) for the inter-annotator and intra-annotation agreement studies reported here. The process in short was to:

1. Use the pre-annotated citation instances to identify annotation areas – i.e. paragraphs that contain at least one citation name. Direct quotes and lists were treated as a part of the same paragraph.
2. Label each sentence in each annotation area as one of *fact*, *principle* or *neither*, following the annotation guidelines.

3.2 Results

Table 1 shows the distribution of categories in the evaluation set of 10 reports. It shows that Annotator 2, who does not have legal training, is more conservative in identifying facts and inferences than Annotator 1, who has had legal training.

The results of the inter-annotator agreement study are as follows: κ=0.65[5] (% Agreement=83.7). The intra-annotator agreement study showed that Annotator 1 (when annotating the same set of 10 reports three months apart in time) was extremely consistent: κ=0.95 (% Agreement=97.3).

[5] κ, the predominant agreement measure used in natural language processing research [5], corrects raw agreement $P(A)$ for agreement by chance $P(E)$: $\kappa = \frac{P(A)-P(E)}{1-P(E)}$

Annotator 1 proceeded to create a gold corpus of 50 reports which was used for training a machine classifier, as described next.

4 Automated annotation study

The methodology used for machine annotation employed classification of the annotation units with a Naive Bayesian Multinomial Classifier based on a set of selected features described below.

4.1 Method

The task of features selection focused on identifying the features that can help in classifying sentences. The following features were selected for extraction from the dataset:

- Unigrams
- Dependency pairs
- Length of the sentence
- Position in the text
- Part of speech tags
- Insent – a feature which indicates whether there is a citation instance in the sentence.
- Inpara – a feature which indicates sentences that were placed within annotation areas, so that sentences that were placed outside it could be filtered out.

Unigrams are widely used in text classification tasks. The performance of classifiers relying on bag-of-words approach can however be impeded by the assumption that word order and grammatical relations are not significant. To address the limitations researchers often complement unigrams by features that can capture dependencies between words. Dependency pairs derived using the Stanford Parser [7] were used to complement unigrams, creating word pairs that are grammatically linked rather than simply collocated like n-grams. Dependency features have previously been shown to be difficult to beat for a variety of text classifications tasks such as sentiment analysis [17] and stance classification [14, 21].

Part of speech tags were selected as a feature for a number of reasons. Firstly, it was expected that modal verbs and verb tense may help to classify the annotation units. Sentences that introduce facts are most often presented in the Past Indefinite tense. For example:

> The contract contained a general condition that in relation to any financial or other conditions either party could at any time before the condition was fulfilled or waived avoid the contract by giving notice.

Secondly, both epistemic and deontic modal qualifiers that use modal verbs are common in sentences containing legal principles, for example:

> It is a question which must depend on the circumstances of each case, and mainly on two circumstances, as indicating the intention, viz., the degree of annexation and the object of the annexation.

4.2 Results

Tables 2–3 report the classification performance of the Naive Bayes Multinomial classifier from the Weka toolkit [13]. The accuracy of the classifier is similar to that of the Annotator 2, who had no legal

	Annotator 1 (original annotation)	Annotator 2 (inter-annotator study)	Annotator 1 (intra-annotator study)
Principles	266 (32%)	211 (26%)	258 (31%)
Facts	56 (7%)	20 (2%)	54 (7%)
Neither	499 (61%)	590 (72%)	509 (62%)

Table 1. Distribution of categories

training in the manual study. This suggests that to the extent such annotations can be carried out based on linguistic principles alone, automated annotation can be performed to the same standard as manual annotation.

	Precision	Recall	F-Measure
Principles	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884

Number of Sentences	2659
Accuracy	0.85
κ	0.72

Table 2. Per category and aggregated statistics for automatic classifier

Machine/Human:	Principles	Facts	Neither
Principles	646	5	160
Facts	4	198	41
Neither	135	38	1432

Table 3. Confusion Matrix

5 Conclusions

An overall analysis suggests that the machine annotation experiment has returned good classification results with Naive Bayesian Multinomial classifier identifying 85% of instances correctly and achieving Kappa equal 0.72. Good combinations of precision and recall have been achieved for all categories (rounding): 82% precision and 80% recall (principles), 82% precision and 81% recall (facts), and 87% precision and 89% recall (neither). Such positive results suggest that the methodology employed as a part of this experiment can provide a suitable basis for further work.

This is a preliminary work on automatic identification of legal principles and facts that are associated with a case citation. To productively deploy a system, further development of a larger and more complex corpus would need to be done. Furthermore, tools to facilitate web-based access to the annotated statements would have to be designed. Such tools would, for example, allow a legal practitioner to not only search, say in Google, for citations mentioned in a case, but also the associated legal principles and facts, providing deep access to and insight into the development of the law. It would also offer the opportunity to access the law directly rather than via the edited and structured materials made available by legal service providers. Finally, we have only addressed accessing cited legal principles and facts, which is distinct from ranking and relating precedents, i.e. Shepardisation. The approach developed here offers some of the source material that could then be used to automate Shepardisation as well as to evaluate given citation analyses.

REFERENCES

[1] Kevin D Ashley and Vern R Walker, 'Toward constructing evidence-based legal arguments using legal decision documents and machine learning', in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pp. 176–180. ACM, (2013).

[2] Awais Athar and Simone Teufel, 'Context-enhanced citation sentiment detection', in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 597–601. Association for Computational Linguistics, (2012).

[3] K Branting, 'Four challenges for a computational model of legal precedent', *THINK (Journal of the Institute for Language Technology and Artificial Intelligence)*, **3**, 62–69, (1994).

[4] Virginia Cano, 'Citation behavior: Classification, utility, and location', *Journal of the American Society for Information Science*, **40**(4), 284, (1989).

[5] Jean Carletta, 'Assessing agreement on classification tasks: The kappa statistic', *Computational Linguistics*, **22**(2), 249–254, (1996).

[6] Blaise Cronin, 'Norms and functions in citation: The view of journal editors and referees in psychology', *Social Science Information Studies*, **2**(2), 65–77, (1982).

[7] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al., 'Generating typed dependency parses from phrase structure parses', in *Proceedings of LREC*, volume 6, pp. 449–454, (2006).

[8] Atefeh Farzindar and Guy Lapalme, 'Legal text summarization by exploration of the thematic structures and argumentative roles', in *Text Summarization Branches Out Workshop held in conjunction with ACL*, pp. 27–34, (2004).

[9] Filippo Galgani, Paul Compton, and Achim Hoffmann, 'Lexa: Building knowledge bases for automatic legal citation classification', *Expert Systems with Applications*, **42**(17), 6391–6407, (2015).

[10] Eugene Garfield, 'Citation indexes for science', *Science*, **122**, 108–111, (1955).

[11] Kent Greenawalt, 'Interpretation and judgment', *Yale Journal of Law & the Humanities*, **9**(2), 5, (2013).

[12] Ben Hachey and Claire Grover, 'Extractive summarisation of legal texts', *Artificial Intelligence and Law*, **14**(4), 305–345, (2006).

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, 'The weka data mining software: an update', *ACM SIGKDD explorations newsletter*, **11**(1), 10–18, (2009).

[14] Kazi Saidul Hasan and Vincent Ng, 'Why are you taking this stance? identifying and classifying reasons in ideological debates', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 751–762, Doha, Qatar, (October 2014). Association for Computational Linguistics.

[15] Timothy L Humphrey, Xin Allan Lu, Afsar Parhizgar, Salahuddin Ahmed, James S Wiltshire Jr, John T Morelock, Joseph P Harmon, Spiro G Collias, and Paul Zhang. Automated system and method for generating reasons that a court case is cited, February 15 2005. US Patent 6,856,988.

[16] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher, 'Information extraction from case law and retrieval of prior cases', *Artificial Intelligence*, **150**(1), 239–290, (2003).

[17] Mahesh Joshi and Carolyn Penstein-Rosé, 'Generalizing dependency features for opinion mining', in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 313–316. Association for Computational Linguistics, (2009).

[18] Florian Kuhn, 'A description language for content zones of german court decisions', in *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts*, pp. 1–7, (2010).

[19] Elizabeth A Leicht, Gavin Clarkson, Kerby Shedden, and Mark EJ Newman, 'Large-scale structure of time evolving citation networks', *The European Physical Journal B*, **59**(1), 75–83, (2007).

[20] Yonatan Lupu and Erik Voeten, 'Precedent in international courts: A network analysis of case citations by the european court of human rights', *British Journal of Political Science*, **42**(02), 413–439, (2012).

[21] Angrosh Mandya, Advaith Siddharthan, and Adam Wyner, 'Scrutable feature sets for stance classification', in *Proceedings of the 3rd Work-*

shop on Argument Mining, ACL 2016, Berlin, Germany, (2016). Association for Computational Linguistics.
[22] Henk F Moed, 'Citation analysis of scientific journals and journal impact measures', *Current Science*, **89**(12), (2005).
[23] Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed, 'Automatic detection of arguments in legal texts', in *Proceedings of the 11th international conference on Artificial intelligence and law*, pp. 225–230. ACM, (2007).
[24] Michael J. Moravcsik and Poovanalingan Murugesan, 'Some results on the function and quality of citations', **5**, 88–91, (1975).
[25] Thom Neale, 'Citation analysis of canadian case law', *J. Open Access L.*, **1**, 1, (2013).
[26] Patti Ogden, 'Mastering the lawless science of our law: A story of legal citation indexes', *Law Libr. J.*, **85**, 1, (1993).
[27] José Plug, 'Indicators of obiter dicta. a pragma-dialectical analysis of textual clues for the reconstruction of legal argumentation', *Artificial Intelligence and Law*, **8**(2-3), 189–203, (2000).
[28] Martin Raz, 'Inside precedents: The ratio decidendi and the obiter dicta', *Common L. Rev.*, **3**, 21, (2002).
[29] John Swales, 'Citation analysis and discourse analysis', *Applied linguistics*, **7**(1), 39–56, (1986).
[30] Simone Teufel, Jean Carletta, and Marc Moens, 'An annotation scheme for discourse-level argumentation in research articles', in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 110–117. Association for Computational Linguistics, (1999).
[31] Simone Teufel, Advaith Siddharthan, and Colin Batchelor, 'Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1493–1502. Association for Computational Linguistics, (2009).
[32] Simone Teufel, Advaith Siddharthan, and Dan Tidhar, 'Automatic classification of citation function', in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 103–110. Association for Computational Linguistics, (2006).
[33] Marc van Opijnen, 'Citation analysis and beyond: in search of indicators measuring case law importance.', in *JURIX*, volume 250, pp. 95–104, (2012).
[34] Radboud Winkels, Jelle De Ruyter, and Henryk Kroese, 'Determining authority of dutch case law', *Legal Knowledge and Information Systems*, **235**, 103–112, (2011).
[35] Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. Approaches to text mining arguments from legal cases, semantic processing of legal texts: where the language of law meets the law of language, 2010.
[36] Adam Z Wyner, 'Towards annotating and extracting textual legal case elements', *Informatica e Diritto: special issue on legal ontologies and artificial intelligent techniques*, **19**(1-2), 9–18, (2010).
[37] Paul Zhang and Lavanya Koppaka, 'Semantics-based legal citation network', in *Proceedings of the 11th international conference on Artificial intelligence and law*, pp. 123–130. ACM, (2007).

Reading Agendas Between the Lines, an exercise

Giovanni Sileno and **Alexander Boer** and **Tom van Engers**[1]

Abstract. This work presents elements for an alternative operationalization of monitoring and diagnosis of *multi-agent systems* (MAS). In contrast to traditional accounts of model-based diagnosis, and most proposals concerning non-compliance, our method does not consider any commitment towards the individual unit of agency. Identity is considered to be mostly an attribute to assign responsibility, and not as the only referent that may be source of intentionality. The proposed method requires as input a set of prototypical *agent-roles* known to be relevant for the domain, and an *observation*, i.e. evidence collected by a monitor agent. We elaborate on a concrete example concerning tax frauds in real-estate transactions.

INTRODUCTION

In previous works [2, 3], we have presented a *model-based diagnosis* view on complex social systems as the ones in which public administrations operate. The general framework is intended to support administrative organizations in improving responsiveness and adaptability, enabled by the streamlining of use cases and scenarios of non-compliance in the design cycle and in operations. This paper focuses in particular on the *operationalization* of model-based diagnosis (to be used in operations, and therefore supporting responsiveness) and differs from the previous papers in granularity, as it provides a specific example of implementation. Note that even if we apply the proposed method to identify the occurrence of non-compliance, it may be used in principle for any other pattern that may be of interest for the organization.

The paper is organized as follows. § 1 provides a general introduction to diagnosis, and to what we intend as diagnosis of social systems; § 2 presents an overview on the various literature in AI about model-based diagnosis; § 3 introduces the case study (sale transactions of real-estates), identifying prototypical scenarios of interest; § 4 concerns the actual exercise of operationalization of monitoring and diagnosis, providing insights and directions for future developments.

1 DIAGNOSIS OF SOCIAL SYSTEMS

In general, a diagnostic process is triggered if there is the presumption that a *failure* occurred in the system. However, what counts as a failure depends on the nature and function of system.

In case of a *designed artifact*, the system is generally associated to a set of requirements, and, at least at the moment of production, to an implementation model—a *blue-print*. A *failure* becomes manifest when there is an inconsistency between the form/behaviour that is observed and what is expected from that artifact. The failure may be at the *design level*, when the implementation does not meet the design requirements; or at the *operational level*, when one of the subcomponents has failed, and propagated its failure to the system.

In case of a *social system* (natural or artificial), the internal mechanisms of social participants are unknown and typically inaccessible. For instance, we are not able to fully know what is in the mind of a person, nor how someone's mind actually works (not even our own).[2] Nevertheless, we still *do* apply (when it is relevant to do so) a *theory of mind* to explain and interpret our own or others' behaviour, by referring to notions as beliefs, desires, and intentions. If we assume that the application of this stance is viable, then, when something goes wrong in a social system, i.e. when someone's expectations about the behaviour of someone else are not met, this means that something went wrong at as informational, motivational, or deliberative level of at least one individual.[3] In order to identify the wrong, however, we have to consider the requirements associated to the system. A first filter of discrimination could be obtained by referring to normative directives: prohibitions and obligations correspond respectively to negative and positive requirements. This would be sufficient, if the contextualization of a generic norm in an actual social setting was straightforward. However, as the existence of the legal system shows, this is far from being the case: the *qualification* of actions, conditions, people and the *applicability* of rules build up the core of the *matter of law* debated in courts. Thus, in an *operational setting*, rather than norms, we need to refer to adequate abstractions of cases, making explicit factors and their legal interpretation; in this way, we handle *contextualized normative models* that can be directly used to discriminate correct from faulty behaviour, all while maintaining a legal pluralistic view.[4]

1.1 Deconstructing identity

Current approaches of diagnosis on MAS consider social system components (software agents, robots, or persons) as individual intentional entities, i.e. following an assumption that could be described as "*one body, one mind*" (see references in § 2.1). In contrast, we assume that intentional entities may transcend the individual instances of the agents. In the case of a *combine* (e.g. in sport, when a player makes an agreement with a bidder on the results of a match) or similar schemes, the collective intentional entity that causes and explains the resulting behaviour is placed behind the observable identities.

[1] Leibniz Center for Law, University of Amsterdam, the Netherlands, corresponding author: g.sileno@uva.nl

[2] In the words of Chief Justice Brian (1478): "for the devil himself knows not the thought of man".

[3] This is true also in domains where the law imputes *strict liability*, i.e. where the claimant only need to prove the occurrence of the *tort*, and not of a *fault* (negligence, or unlawful intent) in the agent who performed the tort. In these cases, the law discourages reckless behaviour, pushing the potential defendant to take all possible precautions. In other words, in strict liability law ascribes fault *by default* to the agents making a tort.

[4] This may be useful for practical purposes: a public administration may for instance use dissent opinions of relevant cases to further strengthen its service implementations.

Such an interpretation of intentionality has relations with the notions of coordination, coalition formation, and distributed cognition.[5] In addition to this "*one mind, many bodies*" scenario, we allow that an agent may interleave actions derived by a certain strategy with actions generated for other intents, independents from the first: the "*one body, many minds*" case may apply as well.

1.2 Diagnosis as part of a dual process

Monitoring agents (e.g. tax administrations) are typically continuously invested with a stream of messages (e.g. property transfer declarations) autonomously generated by social participants. Clearly, they would encounter a cognitive overload if they attempted to reconstruct all "stories" behind such messages.

In affinity with Dual Process theories of reasoning, we may distinguish a *shallower*, less expensive but also less accurate mechanism to filter the incoming messages; and a *deeper*, more expensive, and accurate mechanism to analyze the filtered messages, possibly performing further investigative actions. The first, implemented as a *monitoring* task, is designed by settling what is interesting to be monitored, and which are the threshold conditions that identify *alarming* situations. The second, implemented as a *diagnostic* task, is triggered when such (potentially) alarming situation are recognized, and possibly starts specific courses of actions to look for other clues discriminating possible explanations (diagnostic and non-diagnostic). Note that the two tasks are intimately related: they are both constructed using expectations of how things should go, and of how things may go wrong. Furthermore, planning builds upon abilities, which can be reinterpreted as expectations of how things may go performing certain actions in certain conditions. From a practical reasoning point of view, *planning, monitoring and diagnosis are parts functional to a whole, and the practical reasoning of an agency cannot but be disfigured if one of these functions is neglected.* In other words, all effort that a public administration puts into simplifying the operations in the front-office of service provision (e.g. diminishing the evidential burden on the citizen) should be coupled with effort in the back-office in support of institutional maintenance.

1.3 Side effects

The choice of investigative actions requires some attention as well. In the case of physical systems, *measurements* do not necessarily involve a relevant modification of the studied system (at least at a macro-level), and criteria in deciding amongst alternative measuring methods generally concern costs on opportunities. In the case of a social system, this cannot be the only criterion. For instance, if the target component suspects being under observation, he may adopt an *adversarial* or a *diversionary behaviour* protecting him from intention recognition actions (cf. [28]); he may also drop the unlawful intent as a precaution. In this work, we overlook the planning problem for evidence-gathering tasks taking into account these derived behavioural patterns.

2 RELEVANT LITERATURE

Model-based diagnosis is a traditional branch of study of AI (see e.g. [21] for an overview); it has reached maturity in the 1990s, and it has been applied with success in many domains, reaching a production level of technology readiness (see e.g. [7]). In the following, we retrace the main directions of investigation, highlighting where relevant the specificities of our problem domain.

2.0.1 Consistency-based diagnosis

Early approaches in model-based diagnosis used explicit fault models to identify failure modes (see e.g. [13]), but these evolved towards diagnostic systems based on descriptions of correct behaviour only. Practical reasons explain this progress: in the case of electronic devices, manufacturers provide only descriptions of normal, correct behaviour of their components. Failure modes could be computed simply as inconsistencies with the nominal specifications (cf. [26] for a minimal set of faulty components, [14] for multiple faults configurations). This type of diagnosis is usually called *consistency-based diagnosis*. In short, by having models of correct behaviour of the system components and a topological model of their composition and knowing the initial state, we can predict the expected system state via simple deduction. If the observed output is different, we acknowledge a behavioural discrepancy, which triggers the diagnostic process aiming to identify the faulty components. Note that in this case, such components are deemed *faulty* simply because they do not behave according to their nominal specification: the 'negative' characterization is then constructed in duality to the 'positive' one (cf. *negation as failure*). In recent literature, these are also called *weak fault models* (WFM), see e.g. [35]. This approach entails important consequences: in consistency-based diagnosis, all fault models become equivalent, meaning that, from the diagnoser perspective, "a light bulb is equally likely to burn out as to become permanently lit (even if electrically disconnected)" [15].

2.0.2 Abductive diagnosis

Not surprisingly, the approach provided by consistency-based diagnosis is not fit for certain domains. In medicine, for instance, doctors do not study only the normal physiology of human organisms, but also how certain symptoms are associated to diseases; the hypotheses obtained through diagnosis are used particularly to *explain* given symptoms. In other words, 'negative' characterizations—*strong fault models* (SFM)—are asserted in addition to the 'positive' ones (cf. *strong negation*), rather than in duality to them. In the literature, in order to operationalize this approach, several authors have worked on explicitly characterizing the system with faulty models, starting a line of research which led to the definition of (model-based) *abductive diagnosis* (see e.g. [11], [8]).

2.0.3 Type of diagnosis per type of domain

We can sketch two explanations of why certain domains refer to consistency-based diagnosis, and others to the abductive diagnosis. The first explanation is built upon the use of negation. The former approach takes a *closed-world assumption* (CWA) towards the system domain, while the latter considers an *open-world assumption* (OWA), reflecting the strength of knowledge and of control that the diagnoser assumes having. Reasonably, engineering domains prefer the former (everything that does not work as expected is an error), while natural and humanistic domains usually refer to the latter (there may be a *justification* for why things didn't go as expected). The second explanation considers the different practical function for which diagnosis

[5] cf. [17]: "A central claim of the distributed cognition framework is that the proper unit of analysis for cognition should not be set *a priori*, but should be responsive to the nature of the phenomena under study."

is used in the domain. While by applying consistency-based diagnosis we can identify (minimal) sets of components which are deemed to be faulty and that can be substituted for *repair*, in the second type of diagnosis the underlying goal is to diagnose the 'disease' in order to provide the right *remedy* (that often cannot be a substitution). For these reasons, considering the social system domain, it makes sense to deal not only with positive, normal institutional models (e.g. buyer and seller in a sale contract), but also with explicitly faulty ones (e.g. tax evaders).

Despite these differences, however, abductive diagnosis and consistency-based diagnosis have been recognized as two poles of a spectrum of types of diagnosis [10]. In effect, we find contributions extending consistency-based diagnosis with faulty models (e.g. [15]) and abductive diagnosis with models of correct behaviour. In a more principled way, [25] shows that the two types of diagnosis can be unified relying on a *stable model semantics* (the same used in ASP), essentially because it considers the distinction and separate treatment of *strong negation* and *negation as failure*.

2.0.4 Deciding additional investigations

During a diagnostic process, it is normal to consider the possibility of conducting additional investigations (measurements, in the case of electronic devices) in order to conclusively isolate the set of faulty components, or more generally, to reduce the set of hypothetical explanations. For simplicity, we will neglect this aspect in this work; for completeness, however, we highlight two main directions investigated in the literature. The most frequently used approach, first proposed in [15], is to use a *minimum entropy* method to select which measurement to do next: choosing the datum which minimizes the entropy of the candidate after the measurement is equivalent to deciding the source that provides the maximum *information* to the diagnoser (cf. [**?**]). As this method considers only one additional source per step, it is also called *myopic*. The second approach proposes instead *non-myopic* or *lookahead* methods, i.e. deciding multiple steps to be performed at once, see e.g. [**?**]. In principle, this is the way to proceed when we account strategies for collecting information to minimize or control side-effects.

2.1 Diagnosis of Multi-Agent Systems

The association of diagnosis with *multi-agent systems* (MAS) is not very common in the literature, although the number of studies is increasing. In general, contributions alternatively refer to only one of the two natures of MAS, i.e. mechanism of distributed computation or framework for the instantiation of agent-based models. Therefore, on one side, MAS are proposed as a solution to perform diagnosis of (generally non-agent) systems, like in [27, 24]. On the other side, understanding of social failures is expressed as a problem of social coordination—see for instance [20, 19]. Unfortunately, the latter have generally a design-oriented approach, consequently, non-compliance and social failures are seen has a design issue, rather than systemic phenomena, as would be in a "natural" social system. For this reason, they share a perspective similar to works on checking non-compliance at regulatory level, e.g. [16, 18]: system (normative) requirements are literally taken as the reference on which to test compliance of business processes. Unfortunately, in doing this, we are not able to scope behaviours that superficially look compliant, but, for who knows the 'game', they are not.

Using agent-roles instead of roles The idea of using normative sources is related to the *role* construct; agents are usually seen as enacting certain institutional/organizational roles (e.g. [12]), inheriting their normative characterization. An alternative approach, from which this contribution stems out, has been proposed in [3], constructed on *agent-role* models: constructs that include the coordination of roles. The agent-role model share elements with those used in *intention-recognition* studies, and in particular with those based on logic approaches—see [28] for an overview—grown out from traditional AI accounts of story understanding and abduction. However, from a conceptual point of view, the "first principles" we are considering with agent-roles are not simple rules, but knowledge structures building upon practical reasoning constructs [34] and institutional positions [33]. More importantly, agent-roles are defined not only by a *script*, but also by a *topology*. By allowing to have multiple identities distributed on the topology, the agent-role model enable to take into account the existence of *collective agencies*, transcending the individual social participants.

3 CASE STUDY: SWAP SCHEMES IN REAL-ESTATE TRANSACTIONS

In the following section, we will focus on a well-known type of real-estate fraud, of the family of *swap schemes*, and present a few similar prototypical patterns. In a market context, a swap scheme establishes coordinations between dual groupings of buyers and sellers; as these parties are expected to compete within that institutional framework, it essentially undermines the *arm's length* principle of the market. On small economic scale this is not forbidden: e.g. "if you make me pay less for the guitar that your father is selling, I would make you pay less for my brother's motorcycle." However, in real-estate transactions, property transfer taxes apply. The full interaction includes the tax administration, and in these conditions swap schemes become means to reduce the amount of taxes due and, therefore, are not permitted.

3.1 Outline of a database of scenarios

Let us consider a simplified real estate market, with economic actors buying and selling houses of type A and of type B. Property transfer tax is 6% of the sale price, and the buyer and the seller have both nominally the burden to pay it (the actual distribution amongst the parties is however not fixed a priori). Besides the normal sale, we take into account three different scenarios: a swap scheme implementing a real-estate fraud, a hidden payment, and a wrong appraisal.

Example 1 (REAL ESTATE FRAUD, SWAP SCHEME)**.** *X and Y wants to exchange their properties: X owns a real estate of type A; Y owns one of type B, both worth €10 million. Instead of paying €600,000 per each in taxes, they set up reciprocal sales with a nominal price of €5 million, thus dividing the taxes due in half.*

The scheme is illustrated in Fig. 1. The picture highlights two coordination levels:

- an *intentional coordination* level, generally referring to some composition of institutional roles (in our case buyer/seller structures, the dashed boxes in the figure);
- a *scenario coordination* level, responsible of the synchronization of operations between the intentional coordination structures.

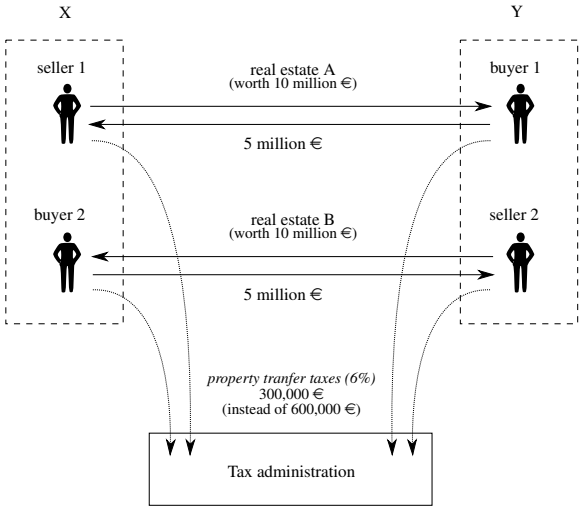


Figure 1. Topology of a real estate fraud based on a swap scheme

The first is the domain of *internal topologies* of agent-roles. The second is the domain of *coupling* configurations of agent-roles, i.e. of *external topologies*, specified as MAS.

The structures enabling coordination (at both levels) may be physical bodies, but also social bodies as natural, informal groupings of people (e.g. father and son), organizations (e.g. employer and employee), etc. It may be anything that suggests a sharing, a *concentration of interests*, or an existence of *stable inter-dependencies*, that may undermine the arm's length principle. At the scenario level, however, the relation is not necessarily as structured as the examples just given. In the case of bribery, for instance, there is typically no other relation between the parties beside a *contingent agreement*. Similarly, a swap-scheme may be performed by two real-estate agencies on a contingent basis.

Example 2 (HIDDEN PAYMENT)**.** *X wants to give €300,000 to Y, and, as Y is also interested in X's house, X sells Y that house, worth €500,000, for €200,000.*

A hidden payment is usually economically advantageous for both parties because property transfer generally has lower taxation than other forms of transfer.

Example 3 (WRONG APPRAISAL)**.** *X needs to sell his house. Not knowing the current prices for the area, he sells the house for €200,000 to Y, while at market price, the building would be worth around €500,000.*

4 OPERATIONALIZATION OF MONITORING AND DIAGNOSIS

In this exercise, we imagine taking the role of the tax administration, with the intent of monitoring the payment of taxes, possibly diagnosing (and also explaining) supposed institutional failures.[6] Note that the tax administration has only a *partial view* of the communications of the parties: in our simplified world, only sale declarations and tax payment receipts.

Types of failures The starting point of the operationalization is to collect the agent-roles of the domain relevant to the tax administration. The first set is given by simple intentional characterizations of *normal institutional roles*, i.e. buyers and sellers paying their taxes. From this, we can construct possible failure modes as violations of role obligations, dealing with representations of *negative events* (negative as they are defined by the failure of expectations concerning events). In this specific example, tax payment may be:

(i) completely missing, as failure to pay *tout court*,
(ii) wrong, as failure to pay the fixed amount of taxes (e.g. 6% of the sale price)
(iii) wrong, as failure to pay the 'right' amount of taxes, in terms of *reasonableness*, i.e. of what could have been expected to be paid to the tax administration for the sale of that property.

The third situation covers the case of swap-schemes or other tax evasion manœuvers; it is evidently more difficult to scope, as it requires an evaluation in terms of the social domain semantics—in this case, of the market pricing rationality. This is the domain in which the agent-role concept makes particularly the difference.

4.1 Monitoring

As we know that certain social participants may be non-compliant, we need to set up an adequate monitoring procedure. A first requirement of adequacy is the possibility of *discriminating* cases of non-compliance from those of compliance. This actually supports a general principle for choosing monitoring targets:

Proposition 1. *Outputs of* contrast *operations between compliant and non-compliant scenarios lead to identifying* events *or* threshold conditions *associated to suspicious transactions.*

The set of discriminating elements is constructed in terms of what is available through the monitoring, i.e. the 'perceptual' system of the agency. If the diagnostic agent is not able to monitor any discriminatory element, then the contrasting principle will not be exploitable and there will be no mean to recognize non-compliance. In our example, as the tax administration has direct access only to sale declarations and tax payment receipts, it is amongst these sources that we have to scope signs of potential failures.

Note that the contrast operation can be implemented thanks to the availability of executable models: by *executing* normal and failure models, we can predict the different traces they would produce, and then contrast them. In principle, however, we could refer directly to the traces. For instance, in medicine, failure modes are usually directly associated to symptoms, without explaining why a certain disease produces these symptoms. In the general case, however, this solution has limitations, as it assumes a relative invariance of the chain of *transmission* going from the source phenomenon to the perceptual system of the observer, which is not granted in a social system. Considering explicitly the underlying behavioural mechanism allows us to deal separately with such 'transmission' component.

We apply the previous principle to the three types of negative events. Case (i) requires the implementation of a *timeout* mechanism that asynchronously triggers the failure. Case (ii) requires a check *synchronously* to the receipt of payment; it can be implemented with a simple operational rule. Case (iii) is more complex: to conclude

[6] It is worth to observe that compliance and non-compliance are qualifications relative to the position of the diagnostic agent in the social system. For instance, in a world of liars, truth-tellers would fail in respect to the social practice of systematically lying.

that a price is reasonable requires us to assess the market price of that property, and to decide what deviation from market price is still acceptable. Let us arbitrarily specify this deviation as 40% of the market price, knowing that statistical methods may suggest more appropriate values. Therefore, the price provided in the sale declaration can be taken as a threshold to consider a certain sale price as *suspicious*. If implemented in Prolog, the qualification rule would look like the following code:

```
suspiciousPrice(Price, Estate, Time) :-
  marketPrice(MarketPrice, Estate, Time),
  Price =< (MarketPrice * 60)/100.

suspiciousSale(Seller, Buyer, Estate, Price, Time) :-
  declaration(sale(Seller, Buyer, Estate, Price, Time)),
  suspiciousPrice(Price, Estate, Time).
```

Clearly, this is a simple case. In general, multiple factors may concur with different weight to increase the suspiciousness of transaction.

In absence of average market price As we confirmed from talking with experts of the tax administration, the practical discrimination used by investigators to discover potential tax frauds is actually built upon comparisons with average market prices. Unfortunately, average market prices are not easy to be access in reality and, when they are, they may be not representative for that specific case.[7] A first solution would then be to refer to domain experts, e.g. appraisal agents, but these externalizations, where available, obviously increase the costs of investigation. A simple way to overcome the problem of assessing the market price of a certain real-estate property is to check the value of the same real-estate in previous sale transactions. In the case of swap schemes, the new owners tend to sell the recently acquired property after a relatively short time, but for a much higher price, even in the presence of relatively stable prices. From an operational point of view, this would correspond simply to a different tracking of the suspiciousness relation.

4.1.1 Diagnosis

When identified, suspicious transactions should trigger a diagnostic process in order to establish *why* the failure occurred. In general, the same 'symptoms' may be associated to diagnostic and non-diagnostic explanations. For instance, going through the known scenarios, a low price in a sale transaction may be due not only to a swap scheme, but also to a hidden payment, or it may simply be due to an error in the appraisal of the estate by the offeror. Interestingly, even if plausible, wrong appraisal is not taken into account by the tax administration. Why? Evidently, this choice is determined by the *strict liability* of these matters[8], but it may be seen as a consequence of a more fundamental issue: the tax administration cannot possibly read the mind of offeror to check the veracity of his declaration. A price that is not 'reasonable' cannot but be interpreted as an *escamotage* of both parties to avoid or reduce the tax burden.

Direct diagnostic mechanism In a simplistic form, direct evidence for a supposed swap-scheme would consist of two sets of buyers and sellers that have performed suspicious sales:

```
actionEvidenceOfSwap(
  sale(Seller1, Buyer1, EstateA, PriceA, Time1),
  sale(Seller2, Buyer2, EstateB, PriceB, Time2)
  ) :-
  suspiciousSale(Seller1, Buyer1, EstateA, PriceA, Time1),
  suspiciousSale(Seller2, Buyer2, EstateB, PriceB, Time2),
  not(EstateA = EstateB),
  not(Seller1 = Seller2), not(Buyer1 = Buyer2).
```

This is however not sufficient: sellers and buyers may have performed these transactions independently, and therefore this evaluation doesn't consider minimal *circumstantial* elements to support a swap-scheme rather than e.g. two hidden payments. In order to overcome this problem, we have to take into account explicitly a *relatedness* condition.

```
actionAndCircumstantialEvidenceOfSwap(
  sale(Seller1, Buyer1, EstateA, PriceA, Time1),
  sale(Seller2, Buyer2, EstateB, PriceB, Time2)
  ) :-
  actionEvidenceOfSwap(
    sale(Seller1, Buyer1, EstateA, PriceA, Time1),
    sale(Seller2, Buyer2, EstateB, PriceB, Time2)
  ),
  relatedTo(Seller1, SharedStructure1),
  relatedTo(Buyer2, SharedStructure1),
  relatedTo(Seller2, SharedStructure2),
  relatedTo(Buyer1, SharedStructure2).
```

An example of *relatedness* condition between buyer and seller may be, for instance, their participation in a common social structure (family, company, etc.), that may place its members outside the arm's length principle of the market. This condition acknowledges *potential* intentional coordination, i.e. a plausible concentration of *interests* that makes the transaction definitively suspect.[9]

The existence of a coordination structure at the scenario level, i.e. between such shared structures, would be additional evidence, but it is not necessary, as the scheme may be performed on a contingent basis (§ 3.1). Interestingly, the 'hidden payment' case turns out to be a minimal version of a swap-scheme:

```
actionAndCircumstantialEvidenceOfHiddenPayment(
  sale(Seller, Buyer, Estate, Price, Time)
  ) :-
  suspiciousSale(Seller, Buyer, Estate, Price, Time),
  relatedTo(Seller, SharedStructure),
  relatedTo(Buyer, SharedStructure).
```

By extension, we could imagine swap-schemes implemented through *networks* of buyer and sellers. This would be, for instance, a simple diagnostic test for swap-schemes performed on three-node networks:

```
actionAndCircumstantialEvidenceOf3Swap(
  sale(Seller1, Buyer1, EstateA, PriceA, Time1),
  sale(Seller2, Buyer2, EstateB, PriceB, Time2))
  sale(Seller3, Buyer3, EstateC, PriceC, Time3)
  ) :-
  suspiciousSale(Seller1, Buyer1, Estate1, PriceA, Time1),
  suspiciousSale(Seller2, Buyer2, Estate2, PriceB, Time2),
  suspiciousSale(Seller3, Buyer3, Estate3, PriceC, Time3),
  not(EstateA = EstateB),
  not(Seller1 = Seller2), not(Buyer1 = Buyer2),
  not(EstateB = EstateC),
  not(Seller2 = Seller3), not(Buyer2 = Buyer3),
  not(EstateA = EstateC),
  not(Seller1 = Seller3), not(Buyer1 = Buyer3),
  relatedTo(Seller1, SharedStructure1),
  relatedTo(Buyer3, SharedStructure1),
  relatedTo(Seller2, SharedStructure2),
  relatedTo(Buyer1, SharedStructure2),
  relatedTo(Seller3, SharedStructure3),
  relatedTo(Buyer2, SharedStructure3).
```

The inclusion of a third element breaks the direct connection between the initial parties, but the code makes explicit the pattern that can be extended by induction. More formally:

[7] On the one hand, prices of real estate properties in public offers often do not correspond to the actual prices of sale. On the other hand, the heterogeneity of real estate properties, the imperfect alignment between cadastral information and real situations, the dynamics of value associated to neighbourhoods and other relevant factors make it difficult to consider as reliable the application of average measures on actual cases.

[8] See note 3.

[9] This is evidently similar to the issue of *conflict of interest*: a person in power may be in a situation in which his discretion to reach the primary intents defined by his role may be biased towards the achievement of other intents.

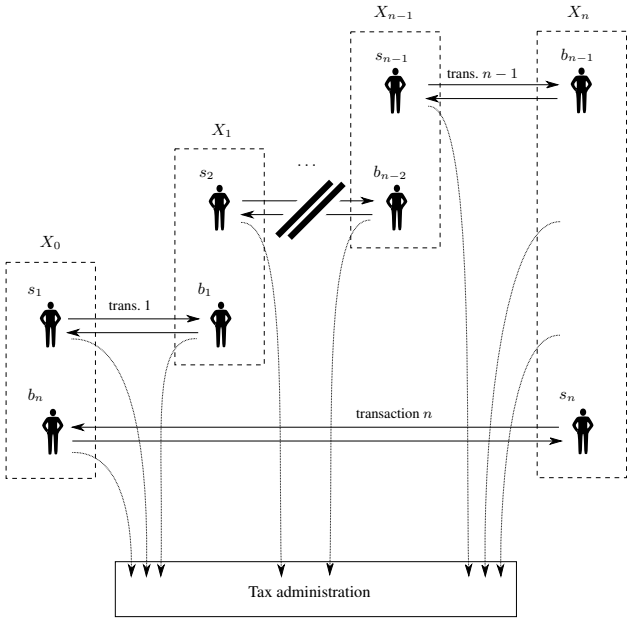


Figure 2. Swap scheme with n nodes.

Definition 1 (GENERALIZED SWAP-SCHEME THROUGH SALES)**.** Given n sale transactions, naming b_i and s_i respectively the buyer and the seller of a transaction i, a swap scheme holds if the following relatedness relations are established:

- between s_1 and b_n (named X_0)
- with $0 < i \leq n$, between s_i and b_{i-1} (named X_i)

The associated topology is illustrated in Fig. 2. It would certainly be interesting to evaluate mechanisms like this on data sets such as those released with the so-called Panama papers.

4.2 Improving the reasoning mechanism

The diagnostic mechanism proposed here leverages the advantages of *backward chaining* given by Prolog, i.e. of reasoning opportunistically in order to reach a conclusion about a certain epistemic goal. In a way, this is an opposite solution than the operationalization we proposed in *explanation-based argumentation* (EBA) [31], based on ASP, where factors brought by the observation are used to allocate *all* possible scenarios. On the other hand, it suffers from two important limitations. First, it relies on a *closed-world assumption* (CWA), i.e. *negation as failure* is automatically interpreted as *strong negation*. Second, it requires an explicit query to trigger the inferential process, but, in a practical setting, the monitoring and diagnostic process should be reactive to the reception of new observations. Therefore, a more plausible monitoring mechanism should look like the following *event-condition-action* (ECA) rule:

(E) when you receive a declaration,
(C) if it is suspicious,
(A) trigger the diagnostic process.

Third, the diagnostic process should consider the whole family of scenarios that are associated to that 'symptom', and should consider that there may be *missing information*. One way to proceed in this respect is to integrate a solution similar to EBA, i.e. of generating at need potential scenarios. Relevant known facts are used to fill fit scenarios belonging to this family, pruning impossible (according to logic constraints), or implausible (according to prior commitments) ones. Note that this family can be compiled *offline*, as much as the discriminatory power of the different factors allow. This information may be used to lead the investigation steps to be acted upon in real-time.

In this scenario, the procedural aspect was not essential, but in general, it may be.In related works, for instance, we built our models using (extensions of) Petri net [30, 32]. Petri net can be mapped to logic programming using for instance Event Calculus [29] or similar techniques; this can be related to *composite event recognition* approaches (e.g. [1]) suggest the use of intermediate caching techniques to improve the search. Another solution would be to instead maintain the process notation, and compute fitness decomposing the family of scenario in a hierarchy of *single-entry-single-exit* (SESE) components (e.g. [23]).

4.2.1 Computational complexity

Model-based diagnosis (MBD) is known to be a hard computational problem, namely exponential to the number of components of the diagnosed systems (see e.g. [4]). For this reason, diagnostic algorithms traditionally focus on minimal diagnoses, i.e. of minimal cardinality (involving minimal subset of faulty components), an approach that is also known as the *principle of parsimony* [26]. This principle is not directly applicable to our framework, as the system components are not agent-players, but agent-roles enacted by agent-players; each component is therefore 'invisible' to the observation, and can be tracked only as a mechanism involving individual elements.

Fortunately, it has been shown that the exponential increase of computational burden may still be reduced using a mixture of decomposition techniques and statistical information. In this chapter, we have overlooked this problem, as we focused on justifying the proposed method providing a working example of an application. We can, however, trace next directions to investigate. As we said in the previous section, the family of scenarios associated to a certain alarming event is known in advance. Therefore, some knowledge compilation techniques may produce important advantages, deriving heuristic knowledge for heuristic problem-solvers, without restarting from first principles (e.g. [5, 9]). Statistical information may instead be used to focus only on a limited set of most probable *leading* hypothesis [15]. It has been also suggested to control complexity by using hierarchical models, i.e. models with different levels of abstraction [22, 6, 35]. This is in principle directly possible with agent-roles. All these aspects remain to be investigated.

5 CONCLUSION

As already stated in the title, this paper is meant to describe an exercise of computational implementation, targeting a specific problem, exploiting part of the conceptual framework presented in previous works [2, 3]. For reasons of opportunity, we neglected many other practical and theoretical aspects that have been investigated in parallel, and that should be taken into account to get the full picture. For instance, about the *representation* of agent-roles, we have identified in *positions* the fundamental components, defined respectively towards another party for normative functions, in the tradition of Hohfeld's analytic framework [33], and towards the environment for practical reasoning purposes [34]. We have investigated the *acquisition* of agent-roles starting from UML-like diagrams [30] and from

interpretations of narratives [32]. In these works we worked with (extensions of) Petri nets, also in order to set a natural convergence to the usual notation used for business process models.

On the other hand, this simplification allowed to appreciate instead the problems of settling a real-time model-based diagnosis activity in operations. It is easy to imagine further developments from the insights gained from this exercise. We will just name a few of them: a formalization of the *contrast* operation; the 'compilation' of the collected scenarios in knowledge bases optimized for monitoring and for diagnosis; the interface of EBA with backward-chaining, in order to take into account competing scenarios and the possibility of missing information; the possibility of composing multiple scenarios via planning, taking into account diversional behaviours (this would not be possible with diagnostic systems not relying on models); an investigation on the resulting computational complexity.

REFERENCES

[1] Alexander Artikis, Marek Sergot, and Georgios Paliouras, 'An Event Calculus for Event Recognition', *IEEE Transactions on Knowledge and Data Engineering*, **27**(4), 895–908, (2015).

[2] Alexander Boer and Tom van Engers, 'An agent-based legal knowledge acquisition methodology for agile public administration', in *Proceedings of the 13th International Conference on Artificial Intelligence and Law - ICAIL '11*, pp. 171–180, New York, (2011). ACM Press.

[3] Alexander Boer and Tom van Engers, 'Diagnosis of Multi-Agent Systems and Its Application to Public Administration', in *Business Information Systems Workshops*, volume 97 of *Lecture Notes in Business Information Processing*, pp. 258–269. Springer, (2011).

[4] Tom Bylander, Dean Allemang, Michael C. Tanner, and John R. Josephson, 'The computational complexity of abduction', *Artificial Intelligence*, **49**, 25–60, (1991).

[5] B. Chandrasekaran and Sanjay Mittal. Deep versus compiled knowledge approaches to diagnostic problem-solving, 1983.

[6] Luca Chittaro and Roberto Ranon, 'Hierarchical model-based diagnosis based on structural abstraction', *Artificial Intelligence*, **155**(1-2), 147–182, (may 2004).

[7] Luca Console and Oskar Dressier, 'Model-based diagnosis in the real world: Lessons learned and challenges remaining', *IJCAI International Joint Conference on Artificial Intelligence*, **2**, 1393–1400, (1999).

[8] Luca Console, Daniele Theseider Dupré, and Pietro Torasso, 'A theory of diagnosis for incomplete causal models', *Proceedings 11th International Joint Conference on Artificial Intelligence*, 1311–1317, (1989).

[9] Luca Console, Luigi Portinale, and Daniele Theseider Dupré, 'Using compiled knowledge to guide and focus abductive diagnosis', *IEEE Transactions on Knowledge and Data Engineering*, **8**(5), 690–706, (1996).

[10] Luca Console and Pietro Torasso, 'A spectrum of logical definitions of model-based diagnosis', *Computational Intelligence*, **7**(3), 133–141, (1991).

[11] P. T. Cox and T. Pietrzykowski, 'Causes for Events: Their Computation and Applications', in *Deductive Databases, Planning, Synthesis - 8th International Conference on Automated Deduction*, volume LNCS 230, pp. 608–621, (1986).

[12] Mehdi Dastani, M. Birna van Riemsdijk, Joris Hulstijn, Frank Dignum, and John-Jules Ch. Meyer, 'Enacting and deacting roles in agent programming', *AOSE 2004: Proc. of 5th Int. Workshop on Agent-Oriented Software Engineering*, (2004).

[13] Randall Davis. Diagnostic reasoning based on structure and behavior, 1984.

[14] Johan de Kleer and BC Brian C. Williams, 'Diagnosing multiple faults', *Artificial intelligence*, **32**(1987), 97–130, (1987).

[15] Johan de Kleer and Brian C. Williams, 'Diagnosis with behavioral modes', *International Joint Conference On Artificial Intelligence*, 1324–1330, (1989).

[16] Guido Governatori, 'Business Process Compliance: An Abstract Normative Framework', *it Information Technology*, **55**(6), 231–238, (2013).

[17] Edwin Hutchins, 'Enaction, Imagination, and Insight', in *Enaction: towards a new paradigm in cognitive science*, 425–450, MIT Press, Cambridge, Massachusetts, (2010).

[18] J I E Jiang, Huib Aldewereld, Virginia Dignum, and Yao-hua Tan, 'Compliance Checking of Organizational Interactions', *ACM Transactions on Management Information Systems*, **5**(4), 1–24, (2014).

[19] Özgür Kafal and Paolo Torroni, 'Exception diagnosis in multiagent contract executions', *Annals of Mathematics and Artificial Intelligence*, **64**(1), 73–107, (mar 2012).

[20] Meir Kalech, 'Diagnosis of coordination failures: a matrix-based approach', *Autonomous Agents and Multi-Agent Systems*, **24**, 69–103, (jul 2012).

[21] Peter J.F. Lucas, 'Analysis of notions of diagnosis', *Artificial Intelligence*, **105**(1-2), 295–343, (1998).

[22] Igor Mozetič, 'Hierachical Model-based Diagnosis', *International Journal of Man-Machine Studies*, **35**(3), 329—-362, (1991).

[23] Jorge Munoz-Gama, Josep Carmona, and Wil M P Van Der Aalst, 'Single-Entry Single-Exit decomposed conformance checking', *Information Systems*, **46**, 102–122, (2014).

[24] M. Pipattanasomporn, H. Feroze, and S. Rahman, 'Multi-agent systems in a distributed smart grid: Design and implementation', *2009 IEEE/PES Power Systems Conference and Exposition*, 1–8, (mar 2009).

[25] Chris Preist, Kave Eshghi, and Bruno Bertolino, 'Consistency-based and abductive diagnoses as generalised stable models', *Annals of Mathematics and Artificial Intelligence*, **11**(1-4), 51–74, (1994).

[26] Raymond Reiter, 'A theory of diagnosis from first principles', *Artificial Intelligence*, **32**(1), 57–95, (apr 1987).

[27] Nico Roos, Annette ten Teije, and Cees Witteveen, 'A protocol for multi-agent diagnosis with spatially distributed knowledge', *Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS '03*, 655, (2003).

[28] Fariba Sadri, 'Logic-based approaches to Intention Recognition', in *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, eds., N.-Y. Chong and F. Mastrogiovanni, 346–375, IGI Global, (2012).

[29] Murray Shanahan, 'The event calculus explained', *Artificial intelligence today*, 409–430, (1999).

[30] Giovanni Sileno, Alexander Boer, and Tom van Engers, 'From Inter-Agent to Intra-Agent Representations: Mapping Social Scenarios to Agent-Role Descriptions', in *Proc. 6th Int. Conf. Agents and Artificial Intelligence (ICAART 2014)*, (2014).

[31] Giovanni Sileno, Alexander Boer, and Tom van Engers, 'Implementing Explanation-Based Argumentation using Answer Set Programming', in *11th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2014)*, (2014).

[32] Giovanni Sileno, Alexander Boer, and Tom Van Engers, 'Legal Knowledge Conveyed by Narratives: Towards a Representational Model', *Proceedings of the Workshop on Computational Models of Narrative (CMN 2014)*, 182–191, (2014).

[33] Giovanni Sileno, Alexander Boer, and Tom van Engers, 'On the Interactional Meaning of Fundamental Legal Concepts', *JURIX 2014: 27th Int. Conf. Legal Knowledge and Information Systems*, **FAIA 271**, 39–48, (2014).

[34] Giovanni Sileno, Alexander Boer, and Tom van Engers, 'Commitments, Expectations, Affordances and Susceptibilities: Towards Positional Agent Programming', *PRIMA 2015: 18th Conf. on Principles and Practice of Multi-Agent Systems*, **LNCS 9387**, 687–696, (2015).

[35] Roni Stern, Meir Kalech, and Orel Elimelech, 'Hierarchical Diagnosis in Strong Fault Models', *DX Workshop*, (2014).

The Implementation of Hohfeldian Legal Concepts with Semantic Web Technologies

Pieter Slootweg and **Lloyd Rutledge** and **Lex Wedemeijer** and **Stef Joosten**[1]

Abstract. This research explores how and to what extent Semantic Web techniques can implement Hohfeldian legal concepts. Laws and regulations are forms of rules in natural language. Because laws are objective and formal, they are suitable for specification with formal logic. Hohfeldian legal concepts are an important tool for the explicit creation of normative legal relationships between the parties. The results of this study show that it is possible for legal requirements based on Hohfeldian legal concepts to be expressed with Semantic Web techniques. For the different Hohfeldian legal concepts, we work out a generic solution within a case study. This work shows that global qualification regarding whether or not a particular action is allowed should not only be determined on the basis of the relevant Hohfeldian legal concepts, but also by taking conditional statements into account.

1 INTRODUCTION

The idea of applying logic to laws and regulations is not new. For some time, scientists have explored the possibilities of deriving legal decisions from legal sources, just as with logical deduction, a conclusion is derived from a set of axioms. However, creating requirements compliant with laws and regulations is difficult. This complexity arises because articles of law are sometimes complementary, overlapping and contradictory.

One method for finding a solution for the difficult task of specifying requirements in legal texts is to focus on the legal norms in the text. Deontic logic is an important way of formally describing these legal norms. Hohfeldian legal concepts constitute a further refinement of the concepts of deontic logic [14]. The primary purpose of Hohfeld's work is to make the normative legal relationships between parties explicit. Hofheldian legal concepts are used in different studies for extracting requirements that are compliant with legal texts. Important examples are Production Rule Modeling (PRM) [16] and the Nomos Framework [20].

A relatively new domain for the implementation of legislation is the Semantic Web. The aim of this study is to investigate how and to what extent Semantic Web techniques can be used to model legal texts with Hohfeldian legal concepts. This work also focuses on the modeling of pre- and post-conditions and exceptions within legal text. The case study we use is the Health Insurance Portability and Accountability Act (HIPAA) [22], partly because HIPAA is also used in several other relevant studies.

This research builds on previous research at the Open University in the Netherlands regarding processing legal texts with formal logic. Bos has done research on the implementation of rules with Semantic Web technologies [7]. Lalmohamed implemented Hohfeldian legal concepts with relation algebra [15]. This relation algebra implementation is a reference for modeling the rules in our study. In comparing the Semantic Web with relation algebra, the main concerns are the open and closed world assumptions and negation as failure.

Francesconi investigated the use of Hofheldian legal concepts based on Semantic Web technologies for the semantic annotation of legal texts [10]. The focus of this research is the application of Hohfeldian legal concepts to the normative qualification of several legal cases within the context of a particular law. The empirical research in our study explores how to use Semantic Web techniques to draw normative conclusions with Hohfeldian legal concepts.

We reuse an existing ontology for modeling Hohfeldian legal concepts - the Provision Model - by extending where necessary for our purposes with our new ontology: HohfeldSW. This is to complete missing Hohfeldian legal concepts and to implement normative qualification. In addition, a domain-specific ontology is elaborated: the HIPAA ontology. Requirements were extracted with normative phrase analysis based on the PRM Method. Our implementation also applied some ontology design patterns such as n-ary relations [18] and AgentRole [19]. These ontology design patterns are of added value to the transparency of the implementation.

The results of our empirical study show that it is possible to express legal requirements based on Hohfeldian legal concepts with Semantic Web techniques. The implementation makes the relationship between actors clear, along with the actions they perform, what the legal consequences are, and if they may or may not perform these actions. With our implementation, it is possible to implement generic rules for validating the various legal concepts.

2 RELATED WORK

2.1 Hohfeldian Legal Concepts

Deontic logic is used to analyze the normative structures and normative reasoning that occur in laws [27]. Deontic logic is formal logic used to reason with ideal and actual behavior: what should be the case, or what should be done by those involved [25]. Deontic logic is developed as modal predicate logic with operators for obligation (O), permission (P) and prohibition (F).

The Hohfeldian legal concepts constitute a further refinement of the concepts of deontic logic [14]. With the aid of the Hohfeldian legal concepts, it is possible to derive the most important legal norms from a text. Hohfeld has developed an analytical scheme in which he distinguishes four categories of legal relations between the parties. He also elaborates on legal differences between the different legal positions [5]. In his view there are eight such entities. On one hand, there are Right, Privilege, Power and

[1] Faculty of Science, Management and Technology, Open University in the Netherlands, Heerlen, The Netherlands, email: Lloyd.Rutledge@ou.nl

Immunity. In addition, there are the correlated entities: Duty, No-right, Liability and Disability.

2.2 Implementations of Hohfeld

Hohfeld's study was widely applied and marked the beginning of a systematic approach. However, this was not enough for a formal theory and a base for the implementation of information systems. Allen and Saxon developed the Hohfeldian legal concepts further into a model in which deontic norm structures could be represented: the A-HOHFELD legal concepts [1]. Allen and Saxon showed in their work how the framework of Hohfeldian legal concepts could be used to define a formal language, which makes it possible to precisely analyze a legal act, thus removing ambiguity.

There are several studies where the Hohfeldian legal concepts are used as a tool to specify legal requirements. Well-known examples include the Nomos framework [20] and the PRM (Production Rule Methodology) [16]. Siena and other researchers developed the Nomos framework to support the requirements analyst in drafting requirements that are compliant with legislation [20]. North Carolina State University focused on the use of formal methods to model legislation. Their focus was on modeling legislation and methods to systematically analyze legal texts. This resulted in the PRM. From the perspective of the PRM, relevant legal concepts are inferred from the words that are used in the normative phrases. Each legal concept also has an implied concept. For example, when a person has a right to a notification made by a hospital, it implies a duty for that hospital to send a notification. The added value of Hohfeld's theory is that implicit assumptions and consequences are made explicit.

Francesconi's model is developed for legislative provisions with axioms from RDFS and OWL [10]. His research makes design patterns with OWL-DL techniques to implement the Hohfeldian legal relationships. The outcome of his research is primarily intended to make a useful contribution to the refinement of semantic annotations to legal texts. The focus of our research is the application of Hohfeldian legal concepts to the normative qualification of various legal cases. We explore the feasibility of this within the context of a specific law: HIPAA [22].

2.3 Law and the Semantic Web

There is much research on the implementation of legislation with Semantic Web technologies. In particular, research on legal ontologies combined with the extraction of semantic standards based on Natural Language Processing (NLP) has given a strong impetus to the modeling of legal concepts [9]. Benjamins has developed a wide variety of ontologies with a wide variety of applications [3]. One demonstration of the importance of legal ontologies is the missing link between AI & Law and Legal Theory [23]. Ontologies for the legal domain are useful in applications such as organizing and structuring information, semantic indexing, integration, reasoning and problem solving.

This research focuses on the application of rules on legal texts, or reasoning and problem solving. Ontologies can thereby be used as a terminology part of a knowledge database in order to derive assertions from the problem to be solved. The role of an ontology in this situation is the representation of domain knowledge so that an automatic logic-reasoning mechanism can represent problems and possibly generate solutions to these problems. Design choices when constructing an ontology are strongly influenced by the ontology's purpose. How knowledge is structured and formalized in the ontology depends on how it is used by the reasoning logic to draw the desired conclusion. The reasoning context limits its reusability in the ontology. This phenomenon is known as inference bias [24]. Inference bias is unavoidable because no wording is completely neutral.

We now present some concrete examples of research on legal ontologies. Wyner developed an ontology in OWL called Legal Case-Based Reasoning (LCBR) [28]. The Leibniz Institute of Law has done extensive research into the development of ontologies for the legal domain. An important ontology in this case is FOLaw (Functional Ontology for Law) [6]. FOLaw specifies functional dependencies between different types of knowledge that are important for legal reasoning. Although FOLaw is an important source for a number of ontologies and legal reasoning systems in various research projects, it is more an epistemological framework than a core ontology. Another important ontology is LKIF, which consists of a core legal ontology and a legal rule language, which makes it possible to represent legal knowledge in detail and reason about it [12]. Other relevant ontologies include Fundamental Legal Conceptions, A-Hohfeld, Language for Legal Discourse, Frame-Based Ontology of Law, LRI-Core [6] and the Core Legal Ontology [11].

Another important development in this context is LegalRuleML. The Technical Committee of OASIS (Advancing Open Standards for the Information Society) developed a rule interchange language for the legal domain. This makes it possible to structure the content of a legal text into machine-readable format, which can be used as a source for further steps such as control and data exchange, comparison, evaluation and reasoning. An important goal in the development of Legal Rule modeling is to bridge the gap between descriptions of natural language and semantic modeling standards [2]. Another important object is to provide an expressive XML standard for modeling of normative rules which comply with requirements from the legal domain. This makes it possible to introduce a legal reasoning layer on top of the ontology.

There are important similarities between LegalRuleML and SBVR (Semantics of Business Vocabularies and Business Rules). We mention SBVR because this is also an important language for specifying rules with Semantic Web technologies. With SBVR concepts, definitions, rules and facts can be expressed in natural language, similar to LegalRuleML. SBVR involves business rules that may or may not have legal significance. LegalRuleML refers to expressions that have legal significance, in particular legal concepts and processes. Distinctive for LegalRuleML are the possibility of defeasibility and the various possibilities for expressing deontic concepts.

2.4 Semantic Web Ontologies for Law

Our study selected the Provision Model [10]. While this ontology is still in development - only some of the Hohfeldian legal concepts are implemented - it is a good basis for our study. This is substantiated by a number of relevant criteria. The Provision Model is implemented transparently. The Provision Model is not only available as an OWL ontology, but is also explained in the aforementioned publication. One of the objectives of the Provision Model is supporting reasoning by making use of normative rules based on Hofheldian legal concepts. The focus is on the derivation of implicit knowledge from explicitly recorded knowledge. The Provision Model is not focused on a specific legal domain, making the risk of misapplication outside the original context limited. The Provision Model meets the criteria for reusability and extensibility because the ontology is specific enough to be reused and, on the other hand, is not too specific so that reuse is impossible. We choose the Provision Model over LKIF-Core [12] because of the

extents of the ontologies and because Hohfeldian legal concepts are not supported directly by LKIF. However, this ontology is a source of inspiration for qualifying legal standards.

3 CONCEPTUAL MODEL

The implementation of Semantic Web technologies is based on three ontologies. The Provision Model [10], based on Hohfeldian legal concepts, is used as a basis. As an extension of this, we designed our own ontology: HohfeldSW. We also developed a domain-specific ontology in OWL, based on the HIPAA Privacy Rule. In the implementation also a number of ontology design patterns are used: AgentRole and n-ary Relations.

3.1 Provision Model

According to Biagioli, legislation can be viewed as a set of 'provisions' (rules) based on speech acts, or more specifically, sentences to which meaning is assigned [4]. A legal text can be viewed from two perspectives on this basis:

1. *Structural or formal perspective*. This is consistent with the traditional classification of a legal text into chapters, articles, and paragraphs.
2. *Semantic perspective*. This is a specific representation of the legal text on the basis of the essential meaning of this text. A possible description can be given in terms of legislative provisions.

From these points of view, components of the legal text are, on one hand, sentences, paragraphs or articles, and on the other hand, provisions, focusing on the semantics. The focus in this study is on the latter. The Provision Model created a division between provision types and related attributes. Examples of types of provision are familiar terms as Duty, Right, Power and Liability. Examples of attributes are Bearer and Counterpart.

In the Provision Model, provision types are divided into two main categories: Rules, and Rules on Rules [10]. The rules of the underlying legal concepts are divided into constitutive and regulatory rules. Rules on rules involve different types of amendments to rules.

The Provision Model extends the standard Hohfeldian legal concepts by making a distinction between implicit and explicit provisions. This comes from the observation that sometimes legal texts mention legal concepts explicitly, but not related correlative legal concepts. For example, a text may explicitly mention a Duty but not a Right. In fact, in a different view of the duty itself, the rollers Bearer and Counterpart can be swapped. An OWL disjoint prevents a concept like Right from being both implicit and explicit.

3.2 HohfeldSW Ontology

The HohfeldSW ontology is our extension on the Provision Model. It introduces a few Hohfeldian legal concepts that are missing in the Provision Model: Privilege-NoRight and Immunity-Disability. Also, SWRL rules have been added for the validation of combinations of pairs Hohfeldian legal concept. We also introduced the concept of qualification. One of the main tasks within the legal domain is applying a particular law in a particular case. It must be established whether or not a particular case is allowed based on the relevant legal norms implemented in the system.

For the cases in which one of the concepts from a specific legal concept Hohfeldian pair is missing, we will have to evaluate whether a particular action is compliant with HIPAA. We implement these cases with SPARQL. We have also integrated the AgentRole [19] pattern in the HohfeldSW ontology.

The AgentRole pattern lets us make claims about the roles of agents without affecting the agents that fulfill these roles. In the HohfeldSW ontology, a stakeholder (agent) plays the role of both Actor and Counterpart. These roles can be coupled via the hasRole object property to a specific individual. The AgentRole pattern is applied to the roles that occur within the HIPAA ontology, such as Covered Entity, Government and Person.

3.3 HIPAA Ontology

The concepts in the HIPAA ontology are filled based on a normative phrase analysis for part of the HIPAA Privacy Rule, based on PRM. Each generic HIPAA Action is elaborated in the form of a conjunction of conditions, which together provide a description of the situation that is associated with that specific HIPAA Action.

In this study, each phrase has a normative Actor (Bearer), a Counterpart, an Action and an Object. Any Action from the HIPAA is linked to Hohfeld legal concept of the Provision Model of HohfeldSW. This is possible because for each legal concept of the Provision Model / HohfeldSW a related "hasBearer 'and' hasCounterpart" object property is available.

In line with research at the Leibniz Center for Law, a norm can be defined as a set of conditions in conjunctive normal form [26]. The norm that a Covered Entity has a privilege to use private health information (PHI) can be defined as follows:

$$N \equiv \text{Use\_private\_health\_information\_privilege} \wedge \exists \text{hasExplicitPrivilegeBearer} \wedge \exists \text{hasExplicitPrivilegeCounterpart} \wedge \exists \text{hasPrivilegeObject}$$

This condition is met in the following situation:

{ *Individual_perform_use_PHI*:
Use_private_health_information_privilege,
Freds' Hospital: **CoveredEntity**, *Fred*: **Person**,
Individual_PHI_for_Use: **Private_health_information_for_using**,
Individual_perform_use_PHI **hasExplicitPrivilegeBearer**
FredsHospital, *Individual_perform_use_PHI*
hasExplicitPrivilegeCounterpart *Fred*,
Individual_perform_use_PHI
hasPrivilegeObject *Individual_PHI_for_Use* }

A normative phrase is identified in HIPAA ontology with a unique legal source identifier based on the related article of the HIPAA Privacy Rule. The legal source is coupled by a hasAction / hasActivity object property to the corresponding Action.

3.4 N-Ary Relations Pattern

In Semantic Web languages like RDF and OWL, a property is a binary relation: it is used to link two individuals together or to link an individual to a value. In some situations, however, it is more obvious to use relationships for certain concepts involving an individual to more than one individual or value is linked these are n-ary relations [18]. In the implementation of this study, relationships in which an individual is associated with multiple other individuals occur at different places. An individual from the class Action_Individual is the relevant concept Hohfeld linked to Bearer, a Counterpart and an Object.

As a generic solution, capturing an n-ary relation involves the creation of a new class represented by new properties [18]. Translated to the HIPAA ontology for any HIPAA Action class defines a relationship with a Bearer Counterpart and an object.

4 IMPLEMENTATION OF HOHFELDIAN LEGAL CONCEPTS

Validation of the implementation will take place at the level of individual stakeholders that interact with each other by performing HIPAA-actions, in which one stakeholder has the role of Actor and the other has the Counterpart role (and vice versa). These interactions may result in conflicting situations and non-compliancy. Each establishment of a legal concept Hohfeld pair gives, when relevant, an indication of the level of its implementation.

4.1 Privilege NoRight legal concepts

The Privilege NoRight legal concept is elaborated in the HohfeldSW pattern. SWRL and SPARQL are used for the validation. With OWL, it is possible to infer implicit knowledge from explicit knowledge which is present in the model. This is consistent with the derivation of an implicit legal concept from the correlated explicit legal concept. Table 1 shows an overview of relevant OWL- DL axioms.

The rdfs:subPropertyOf axiom is used to implement a logical implication: if there is a ExplicitNoRightCounterpart then a NoRightCounterpart is implied. An object property can be linked to a certain domain: in this case, hasNoRightCounterpart is linked to the NoRight class. In this way, a Bearer can be coupled to the relevant legal concept class. The classes and object properties for the other legal concept pairs are implemented in a similar way.

Actor Fred's Hospital has the freedom (Privilege) to use Fred's private health information (PHI). Actor Fred has no right to do something about it (No-Right). When Fred tries nonetheless to prohibit the use of PHI, then an infringement occurs. Validation is effected by means of two scenarios. Scenario 1 assumes both a 'Privilege' as a 'No-Right'. Scenario 2 is only the 'Privilege' action 'use PHI'.

Table 1. OWF axioms for NoRight

RDFS/OWL	Example
owl:subClassOf	ExplicitNoRight ⊆ NoRight ImplictNoRight ⊆ NoRight
owl:EquivalentClass	ExplicitNoRight ≡ ImplicitPrivilege ImplicitNoRight ≡ ExplicitPrivilege
rdfs:subPropertyOf	hasExplicitNoRightCounterpart ⊆ hasNoRightCounterpart hasImplicitNoRightCounterpart ⊆ hasNoRightCounterpart
owl:equivalentProperty	hasExplicitNoRightCounterpart ≡ hasImplicitPrivilegeCounterpart hasImplicitNoRightCounterpart ≡ hasExplicitPrivilegeCounterpart

4.1.1 Scenario 1: SWRL

Step 1: Fred prohibits the use of PHI by FredsHospital (NoRight)
Step 2: Fred's Hospital uses PHI Fred (Privilege)

This is documented in the following triples:

```
Fred performProhibitUsePHI Individual_perform_prohibit_use_PHI .
Fred interactWith FredsHospital .
FredsHospital performUsePHI Individual_perform_use_PHI .
```

A generic SWRL rule validates that there is both a privilege and a NoRight same actors (with opposing roles):

```
NoRight(?x), hasNoRightBearer(?x, ?a), hasNoRightCounterpart(?x, ?b), hasNoRightObject(?x, ?o), Privilege(?y), hasPrivilegeBearer(?y, ?b), hasPrivilegeCounterpart(?y, ?a), hasPrivilegeObject(?y, ?o) -> PrivilegeNoRightDisallowed(?y)
```

Comparison with the validation rules of the relation algebra implementation by Lalmohamed [15] helps to identify the related concepts in the Semantic Web implementation. A SWRL rule is needed for testing if both correlative legal concepts are present for an individual (intersection).

The implementation of Scenario 1 gives an individual within PrivilegeNoRightDisallowed. If in addition to a 'Privilege', a 'No-Right' action occurs. This constitutes a breach of privilege. Figure 1 shows this breach being displayed by the Semantic Web tool Protégé. In general, it is possible to check the presence of the two correlative Hohfeldian legal concepts by means of a rule-SWRL.

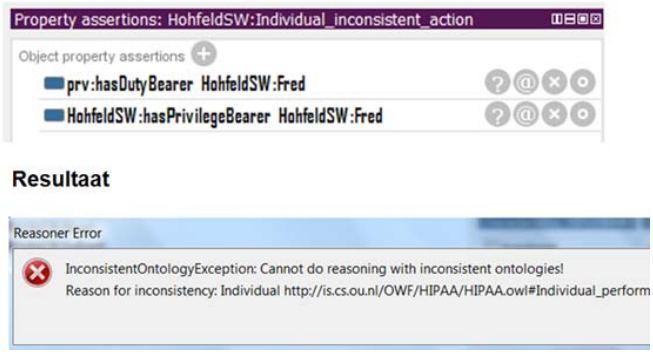


Figure 1. An action cannot be both a Duty and Privilege

4.1.2 Scenario 2: SPARQL vs. closed world-assumption

Step 1: Actor Fred's Hospital uses PHI Fred

This is documented in the following triples:

```
FredsHospital performUsePHI Individual_perform_use_PHI .
FredsHospital interactWith Fred .
```

RDFS, OWL and SWRL cannot establish whether a particular situation does not occur because of the open world assumption. In order to establish that there is an explicit Privilege action, but no corresponding Right-action, the following SPARQL code can be used:

```
INSERT {?ActiePrivilegeAllowed a
  HohfeldSW:PrivilegeNoRighAllowed}
WHERE
{?ActiePrivilegeAllowed a HohfeldSW:ExplicitPrivilege .
 ?ActiePrivilegeAllowed HohfeldSW:hasExplicitPrivilegeBearer
   ?CoveredEntity .
 ?ActiePrivilegeAllowed HohfeldSW:hasExplicitPrivilegeCounterpart
   ?Person .
 ?ActiePrivilegeAllowed HohfeldSW:hasPrivilegeObject ?Object
NOT EXISTS {?NoRightActie a HohfeldSW:ImplicitNoRight .
 ?NoRightActie HohfeldSW:hasImplicitNoRightBearer ?Person .
 ?NoRightActie HohfeldSW:hasImplicitNoRightCounterpart
   ?CoveredEntity .
 ?NoRightActie HohfeldSW:hasNoRightObject ?Object }}
```

This scenario provides an individual Individual_perform_use_PHI in the class PrivilegeNoRightAllowed. The SPARQL code is divided into two conjunctive elements: the conditions in respectively the WHERE and the NOT EXISTS clause. The conditions in the WHERE clause determine whether or not there is a valid Privilege action. The NOT EXISTS clause assesses that there is no NoRight action with a Bearer, Counterpart and object related to the Privilege action. The conjunctive part of the WHERE clause is in line with research into HARNESS [8] and formalized in the following way:

GC_1_Where_Cl ≡ Action(?a) ∧ ExplicitPrivilege(?e) ∧
CoveredEntity (?c) ∧ Person (?p) ∧ ObjectOfAction(?o) ∧ a(?a,?e) ∧
hasExplicitPrivilegeBearer(?a, ?c) ∧
hasExplicitPrivilegeBearer(?a,?p) ∧ hasPrivilegeObject (?a,?o)

For the sake of completeness, implicit concepts such as Person and Covered Entity are named explicitly. Note that the conditions are, to a large extent, similar to the body (condition) of SWRL rules. The ability to apply the variables in SPARQL makes it easier and more transparent to specify the conditions in a query.

This SPARQL solution is applicable in a similar way for the other pairs of correlative Hohfeldian legal concept. In our implementation, we use SPARQL on one hand to establish that a particular action does not occur (negation), and one the other hand to draw a conclusion about the classification of the action (inferencing).

The only other way to make a distinction between a potential Prohibit Use PHI action and the fact that a Prohibit Use PHI really is not applicable is to indicate explicitly that this action really does not take place, for example, in the following way:

FredsHospital performNoProhibitUsePHI
Individual_perform_no_prohibit_use_PHI .

In HohfeldSW ontology, a separate class NoRuleAvailable with relevant subclasses (like NoRightNotAvailable) can be created, which can then be used in a SWRL rule for validation in the form of:

NoRightNotAvailable(?x), hasRelatedAction(?x, ?y), Privilege(?y),
hasPrivilegeBearer(?y, ?b), hasPrivilegeCounterpart(?y, ?a),
hasPrivilegeObject(?y, ?o) -> PrivilegeNoRightAllowed(?y)

With the object property hasRelatedAction, the action which does not occur, Individual_perform_no_prohibit_use_PHI, can be linked to the action Individual_perform_Use_PHI. When Individual_perform_no_prohibit_use_PHI is made member of NoRightNotAvailable class, then application of the SWRL rules shows indeed that Use PHI is permitted.

4.2 Right-duty legal concepts

The Right-duty legal concept is part of the Provision Model. Compliancy is determined through the Qualification concept. SWRL and SPARQL are used for the validation. The actor Fred has the right for a notification if his private health information is used by counterpart Fred's Hospital. In addition, actor Fred's Hospital has the duty to send a notification. Two scenarios are used for validation. In Scenario 1, both a Right and Duty action are used. Scenario 2 assumes a Right action and a pre-condition (Use PHI)

4.2.1 Scenario 1: SWRL rule, conditional statement, sequence actions

Step 1: Actor Fred asks for a notification to Fred's Hospital (Right)
Step 2: Actor Fred's Hospital will send a notification Fred (Duty)

This is documented in the following triples:

Fred performRequestNotification
Individual_perform_request_notification .
Fred interactWith FredsHospital .
FredsHospital performSendNotification
Individual_perform_send_notification .

A SWRL rule can determine that there is both a Right- and Duty action for the same stakeholders (opposite roles):

Right(?x), hasRightBearer(?x, ?a), hasRightCounterpart(?x, ?b),
hasRightObject(?x, ?o), Duty(?y), hasDutyBearer(?y, ?b),
hasDutyCounterpart(?y, ?a), hasDutyObject(?y, ?o) ->
RightDutyAllowed(?y)

Implementation of the SWRL rule provides an individual in the class RightDutyAllowed. This is correct from the perspective of reasoning with Hofheldian-legal concepts. Yet this does not provide a satisfactory qualification. Fred's Hospital has only the duty to send a notification when Fred's private health information is actually used. The duty to send a notification is conditionally dependent on the use of private health information. In this study, conditional dependence has been implemented by means of a pre-condition. The pre-condition Use PHI is not fulfilled in this case, resulting in an individual in class PreConditionNotFulfilled.

It is interesting to determine what would be a logical 'total' qualification of both the Right-Duty Hohfeldian legal concept couple as the conditional dependence. The pre-condition we use here is only a pre-condition for the Duty action. As expected, only when the pre-condition is not fulfilled will this have an impact on the final qualification, in which the final classification is different from the classification on the basis of Hohfeldian legal concepts. Table 2 shows this for all combinations of the Right-Duty Hohfeldian legal concepts where the Duty pre-condition is not fulfilled.

Table 2. Qualification Right-Duty with precondition for Duty action

Right	Duty	Right-Duty Qualification	Resulting Qualification
None	None	None	None
Request notification		Disallowed	Allowed
None	Send notification		Disallowed
Request notification		Allowed	Disallowed

It is notable that for both our Semantic Web implementation and Lalmohamed's relation algebra implementation, there is a challenge with respect to the modeling of the sequence of actions. Although the user interface of the relation algebra implementation can specify a sequence of actions, this is inferred entirely from pre-specified Hohfeldian legal action pairs, without taking into account the sequence of related actions. With Semantic Web technologies it is possible to use a Data Property "action time" in combination with numeric comparison in SWRL, to determine the order of the different actions.

4.2.2 Scenario 2: SPARQL query pre-condition

Step 1: Actor Fred asks for a notification to Fred's Hospital (Right)
Step 2: Actor Fred's Hospital uses PHI Fred (Privilege)

This is defined by the following triples:

```
Fred performRequestNotification
Individual_perform_request_notification .
Fred interactWith FredsHospital .
FredsHospital performUsePHI Individual_perform_use_PHI .
```

Also, for the validation of situations in which the Duty-action is missing, it is relevant to take into account the pre-condition. In case the pre-condition of the Duty action is not fulfilled (Fred's PHI is not used), there is also no need for the Duty action. While in this case, in which the PHI of Fred is used, a Duty action is mandatory. The following SPARQL query is developed for this situation:

```
INSERT {HIPAA:Individual_perform_request_notification a
  HohfeldSW:RightDutyDisallowed}
WHERE {  ?RightAction a HIPAA:Request_notification.
  ?Person HIPAA:performAction  ?RightAction .
  ?Person HohfeldSW:interactsWith ?Hospital .
  ?PrivilegeAction a
    HIPAA:Use_private_health_information_privilege .
  ?Hospital HIPAA:performAction ?PrivilegeAction .
  ?Hospital HohfeldSW:interactsWith ?Person
NOT EXISTS { ?DutyAction a HIPAA:Send_notification.
  ?Hospital HIPAA:performAction ?DutyAction.
  ?Hospital HohfeldSW:interactsWith ?Person }}
```

This results in a qualification RightDutyDisallowed for the same stakeholders (albeit in other role). This is applicable to both a Request notification and a Use PHI, but not a Send notification.

4.3 Power-Liability legal concepts

The Power Liability legal concept is part of the Provision Model. Compliancy is determined through the Qualification concept. SWRL and SPARQL are used for the validation. In the example here, the actor Fred's Hospital has the power to stop the restriction of private health information. Fred is liable to agree to end the restriction. Agreement with the restriction is in contradiction with the Power of Fred's Hospital. Two scenarios can be distinguished. In scenario 1, there is both a Power and Liability. In scenario 2, there is only Power action.

4.3.1 Scenario 1: SWRL

Step 1: Fred agrees to the restriction of PHI by Fred's Hospital
Step 2: Fred's Hospital eliminates the restriction of Fred's PHI.

This is defined by the following triples:

```
Fred performAgreeToRestrict_Liability
Individual_perform_agree_to_restrict_PHI .
FredsHospital interactWith Fred .
FredsHospital performTerminateRestriction
Individual_perform_terminate_restriction .
```

The implementation is assumed that if there is a Liability action, then it undermines the Power action. The following SWRL rule validates this:

```
Power(?x), hasPowerBearer(?x, ?a), hasPowerCounterpart(?x, ?b),
hasPowerObject(?x, ?o), Liability(?y), hasLiabilityBearer(?y, ?b),
hasLiabilityCounterpart(?y, ?a), hasLiabilityObject(?y, ?o) ->
PowerLiabilityDisallowed(?y)
```

This results in an individual in the class PowerLiabilityDisallowed.

4.3.2 Scenario 2: SPARQL

Step 1: Actor Fred's Hospital eliminates the PHI Fred restriction.

This is defined by the following triples:

```
FredsHospital performTerminateRestriction .
Individual_perform_terminate_restriction .
FredsHospital interactWith Fred .
```

The validation of the missing liability action with the absence of negation of failure in the context of Semantic Web can only be resolved by means of SPARQL. The following generic SPARQL query validates this scenario:

```
INSERT {?PowerAction a HohfeldSW:PowerLiabilityAllowed }
WHERE {?PowerAction prv:hasPowerBearer ?PowerBearer.
     ?PowerAction prv:hasPowerCounterpart ?PowerCounterpart.
     ?PowerAction prv:hasPowerObject ?PowerLiabilityObject
NOT EXISTS {
      ?LiabilityAction prv:hasLiabilityBearer ?PowerCounterpart .
      ?LiabilityAction prv:hasLiabilityCounterpart ?PowerBearer .
      ?LiabilityAction prv:hasLiabilityObject ?PowerLiabilityObject }}
```

Validation provides an individual in the class PowerLiabilityAllowed.

4.4 Immunity-Disability legal concepts

The Immunity-Disability legal concept is developed in the HohfeldSW ontology. The Immunity-Disability legal concept does not exist in the HIPAA Privacy Rule [15]. For demonstration purposes therefore a fictional normative phrase is developed.

A government 'Governement1' has a disability related to Fred's Hospital to prohibit the use of private health information. Fred's Hospital is immune for actions from the government to ban the use of private health information. The foregoing is validated with two scenarios. In scenario 1, there is both an Immunity action as a Disability action. In scenario 2, there is only an Immunity action. Validation of both scenarios occurs in a similar manner as in the Power Liability legal concept.

4.5 Opposing legal concepts

We apply OWL for validation of opposing legal concepts. A legal concept is opposed if the existence of one action rules out the existence of the other action. If action "*use private health information*" is a privilege then it cannot simultaneously be a duty because a privilege is part of the PrivilegeNoRight relationship, resulting in a different legal relationship between Actor and Counterpart. The classes in the HohfeldSW ontology are explicitly disjoint. This triggers an inconsistency message stating that the rules are contradictory. Table 3 shows the implemented disjoints.

Table 3. Opposing legal concepts

Legal concept	Disjoint With
Right	NoRight
Duty	Privilege
Power	Disability
Liability	Immunity

5 CONCLUSION

The results of this empirical study show that it is indeed possible to express legal requirements based on Hohfeldian legal concepts with Semantic Web technologies. The implementation clarifies the relationship between actors, what actions they perform and what the legal consequences are, and whether they may or may not perform these actions.

To answer the main question, with the focus on 'how and to what extend' we used a hybrid approach. On one hand for certain parts a formal logic approach was used by applying a set of conditions as a conjunctive norm. On the other hand, design principles for ontologies where used, for instance the good practice of reusability. Furthermore design patterns and normative phrase analysis played an important role in the implementation.

This study used an existing ontology as a foundation: the Provision Model. The Provision Model was no ready-made solution, but a good starting point for the implementation of Hohfeldian legal concepts. The Provision Model misses some Hohfeldian legal concepts. In this study a new ontology is developed: HohfeldSW which extends the Provision Model. In addition to legal concepts not available in the Provision Model, HohfeldSW also adds validation rules and classes to qualify legal acts. This implementation also uses ontology design patterns: n-ary relations and AgentRole.

The development of Hohfeldian legal concepts alone is insufficient to model legislation in a realistic way. In practice, laws and regulations have all kinds of dependencies between rules. In order to be able to proceed, it is necessary to model conditional statements. This is done in the form of pre- and post-conditions and exceptions.

The comparison of the Semantic Web implementation with the relation algebra implementation provides a basis for the level of implementation. The Provision Model itself was able to be implemented at the level of RDFS and OWL. Semantic Web technologies validate correlative legal concept pairs in two ways. Validation of the correlative legal concepts takes place with SWRL if something prevents both legal concepts in the relevant correlative pair, and in other cases with SPARQL because of the open world assumption. However, it is possible to provide a generic solution in all cases. The validation of opposing legal concepts is implemented with a disjoint. In addition, the treatment of pre- and post-conditions and exceptions are implemented with SWRL as SPARQL as well.

In this study, it became clear that the overall qualification about whether a particular action is or is not allowed cannot be determined on the basis of the relevant Hohfeldian legal concepts alone. Conditional statements must be factored in. Finally, it should be noted that although it has been possible to work out generic solutions for drawing conclusions normative and for cross-references, this did not happen entirely at the level of RDFS and OWL.

The source code made for this research is available online[2].

[2] http://is.cs.ou.nl/OWF/index.php5/Hohfeld_with_Semantic_Web

6 ACKNOWLEDGEMENTS

This work was executed as part of the Master's Thesis of Pieter Slootweg. Funding for this work comes from the Faculty of Science, Management and Technology at the Open University in the Netherlands [21].

REFERENCES

[1] Allen, L., & Saxon, C. (1995). Better language, better thought, better communication: the A-Hohfeld language for legal analysis. 219-228. doi:10.1145/222092.222245.

[2] Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., & Wyner, A. (2013). OASIS LegalRuleML. Paper presented at the Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, Rome, Italy.

[3] Benjamins, V. R., Casanovas, P., Breuker, J., & Gangemi, A. (2005). Law and the semantic web, an introduction (Vol. 3369, pp. 1-17). Springer-Verlag, Berlin.

[4] C. Biagioli, "Law making environment: model based system for the formulation, research and diagnosis of legislation", *Artificial Intelligence and Law*. 1996

[5] A.W.F. Boer, "Legal Theory, Sources of Law and the Semantic Web", In: *Frontiers in Artificial Intelligence and Applications*, **195**, 2009.

[6] Breuker, J., & Hoekstra, R. (2004). Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law.

[7] P. Bos, *Bedrijfsregels in verschillende vormen-Een vergelijking op toepasbaarheid tussen SWRL en Relatie algebra bij wetteksten*, Masters Thesis, Open University in the Netherlands, Heerlen, Netherlands, 2013.

[8] Förhécz, A., and Strausz, G. (2009). Legal Assessment Using Conjunctive Queries. IDT, 1.

[9] Francesconi, E., Montemagni, S., Peters, W., & Tiscornia, D. (2010). Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language (Vol. 6036): Springer.

[10] E. Francesconi, 'Semantic model for legal resources: Annotation and reasoning over normative provisions', *Semantic Web Journal*, **7**, 255-265, 2016.

[11] Gangemi, A., & Presutti, V. (2009). Ontology design patterns Handbook on Ontologies (pp. 221-243): Springer.

[12] Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF Core Ontology of Basic Legal Concepts. LOAIT, 321, 43-63.

[13] R.J. Hoekstra, *Ontology Representation: design patterns and ontologies that make sense*. PhD thesis, University of Amsterdam, Amsterdam, 2009.

[14] W.N. Hohfeld, "Fundamental Legal Conceptions as Applied in Judicial Reasoning", *Yale Law Journal*, 23(26(8)), p. 710-770. 1917.

[15] A. Lalmohamed, *Expressing Hohfeldian legal concepts, traceability and ambiguity with a relation algebra-based information system*, Master's thesis, Open University in the Netherlands, Heerlen, Netherlands, 2014.

[16] Maxwell, J. C., & Anton, A. I. (2010). A refined production rule model for aiding in regulatory compliance.

[17] Maxwell, J. C. (2011). A legal cross-references taxonomy for identifying conflicting software requirements.

[18] Noy, N., & Rector, A. *Defining n-ary relations on the Semantic Web*. http://www.w3.org/TR/swbp-naryRelations/. 2006

[19] V. Presutti. *AgentRole ontology pattern*. http://ontologydesignpatterns.org/wiki/index.php?title=Submissions:AgentRole. (2008)

[20] Siena, A., Mylopoulos, J., Perini, A., & Susi, A. (2009). Designing law-compliant software requirements. In e. a. A.F. Laender (Ed.), Conceptual Modeling-ER 2009 (pp. 472-486): Springer.

[21] P. Slootweg, *De implementatie van Hohfeldian legal concepts, ambiguïteit en traceerbaarheid met Semantic Web-technologieën*,

Masters thesis, Open University in the Netherlands, Heerlen, Netherlands, 2016.

[22] U.S. Department of Health & Human Services, *The HIPAA Privacy Rule*, http://www.hhs.gov/hipaa/for-professionals/privacy/index.html.

[23] Valente, A., & Breuker, J. (1994). Ontologies: The missing link between legal theory and AI & law. In H. Prakken, A. J. Muntjewerff, & A. Soeteman (Eds.), Legal Knowledge Based Systems Jurix 1994. Lelystad: Vermande.

[24] Valente, A. (2005). Types and Roles of Legal Ontologies. In V. R. Benjamins, P. Casanovas, J. Breuker, & A. Gangemi (Eds.), Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications (pp. 65-76). Berlin, Heidelberg: Springer Berlin Heidelberg.

[25] T. Van Engers, A Boer, J. Breuker, A. Valente, and R. Winkels, "Ontologies in the legal domain", *Digital Government* (pp. 233-261), Springer. 2008.

[26] Ven, S. v. d., Breuker, J., Hoekstra, R., & Wortel, L. (2008). Automated Legal Assessment in OWL 2. Paper presented at the Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference.

[27] R.J. Wieringa and J.-J.C. Meyer, "Applications of deontic logic in computer science: a concise overview", *Deontic logic in computer science,* M. John-Jules Ch & J. W. Roel (Eds), pp. 17-40, John Wiley and Sons Ltd, 1993.

[28] Wyner, A. (2008). An ontology in OWL for legal case-based reasoning. Artificial Intelligence and Law, 16(4), 361-387.

CALCULEMUS: Towards a Formal Language for the Interpretation of Normative Systems

Robert van Doesburg and **Tijs van der Storm** and **Tom van Engers**[1]

Abstract. In this paper we will introduce a method for creating computational models of normative systems deduced from sources of norms in natural language.

The authors show how to make formal interpretations of normative sources in natural language that result in a computational model, which includes explicit references to all sentences of sources of norms that are considered relevant by the interpreters to constitute a computational model. The models produced can easily be held in sync with these sources.

The method presented is focused on the translation of laws in natural language into an executable computational form that can be easily validated by legal experts that have to decide on the desired interpretation of the source text. The model is tested in a prototype of a reasoner build in a newly developed domain specific language: FLINT. The model is based on Hohfeld's fundamental legal concepts.

1 INTRODUCTION

Organizations that handle thousands or even millions of cases a year depend on a form of computational law to be able to use supporting IT-systems. These organizations are accountable for building and maintaining such systems in compliance to the norms they are submitted to. This is the work of knowledge engineers that use experts' knowledge elicitation processes to incorporate these experts' interpretations of the normative sources of their organizations.

The two primary sources of norms are: legislation, i.e. bills and operational policy documents that all typically describe how generic abstract cases are to be treated, and case decisions in judicial procedures, from which we may learn how a specific individual case is to be treated, and that might have an impact on future cases too.

Knowledge engineers are typically intermediating between the (legal) experts and technical IT staff. They lack a method to formally link the knowledge of the elicited domain experts to the normative sources in natural language that these domain experts use to acquire their knowledge. This is especially problematic in case of changes in normative sources. Organizations need to quickly understand the impact of such changes and adapt their supporting IT-systems accordingly.

[1] Leibniz Center for Law, University of Amsterdam, Netherlands, email: RobertvanDoesburg@uva.nl;
CWI, Netherlands, email: Storm@cwi.nl
Leibniz Center for Law, University of Amsterdam, Netherlands, email: vanEngers@uva.nl

In the early nineties and the first decade of the twenty-first century solutions for this problem where presented [1][16][17], but none of these methods are presently being used on production scale within governmental organizations or industries. In this paper we will shortly describe the difference between our approach and early work. An elaborated overview of the various earlier approaches and the relation to our work will be published as a separate paper, this paper is too short for that exposé.

In this paper we present our approach, called CALCULEMUS, after the ideas of Leibniz who was the first that aimed at solving legal problems by means of calculation. We will demonstrate how it can be applied on actual legal sources with an example from Dutch Immigration Law. The resulting model is expressed in a domain specific language (DSL), FLINT (Formal Language for the Interpretation of Normative Theories). This DSL is specific in so far that it is targeted towards the specific way we express norms. We will illustrate this by giving an example of FLINT expressions.

The CALCULEMUS method and the FLINT prototype result from a co-operation between the Dutch Immigration and Naturalisation Service (IND), the Dutch Tax and Customs Administration (DCTA) and the Leibniz Center for Law. This paper is a report on the progress made on this subject since the NWO Workshop ICT with Industry in December 7-11, 2015 [11].

2 RELATED WORK

Our approach is based on the work of Wesley Newcomb Hohfeld and the fundamental legal concepts he introduced in 1913 [6]. Hohfeld's motive to introduce these legal concepts was his opinion that one of the greatest hindrances to the clear understanding of legal problems is the explicit or tacit assumption that all legal relations may be reduced to "rights" and "duties". Hohfeld proofed this was not the case by describing the ambiguities in the meaning of these concepts and went on to introduce a smallest set of legal conceptions to which, according to him, any and all 'legal quantities' could be reduced.

Hohfeld distinguished four Legal Relations: 'Power-Liability relations' (1), 'Immunity-Disability relations' (2), 'Duty-Claimright relations' (3), and 'Privilege-Noright relations' (4). Some scholars prefer 'Liberty-Noright relations' instead of 'Privilege-Noright relations'. We also use the first term.

The Hohfeldian legal conceptions can only exist in pairs and describe relations between two people, each holding one of the rights in a pair. 'Power-Liability relations' and 'Immunity-Disability relations' are generative: they can generate new 'Legal Relations'. The 'Duty-Claimright relations' and 'Privilege-Noright

relations' are situational: they can only be created and terminated by a generative 'Legal Relation'.

To make our interpretations maximum traceable to the normative sources they are based upon, we strive for isomorphism. Trevor Bench-Capon [1][2] and Van Engers [12][15] are amongst the people that have stressed the importance of creating isomorphism between the formal models that represent sources of law and those sources.

Compared to the method presented in 1991 by Bench-Capon the CALCULEMUS approach is more precise in the explicit notation of references between sentences in normative sources. The fact that we have good and (inter)nationally accepted mechanisms for representing references and standards for identifying building block in sources of law, such as the Dutch Juriconnect standard and the European MetaLex standard [3] helps enormously.

In addition to that in CALCULEMUS a method for the formal interpretation of norms is used that represents the *rules of the game*.

Those rules can be used to actually *play the game*, but making models that describe *games*, i.e. models that include agency, intent and the dynamics of social interaction, are a separate issue. In this paper we will restrict ourselves to making formal interpretations of the *rules of the game*.

The method for interpreting norms used in CALCULEMUS, has similarities with Van Kralingen's norm-framework approach [16]. Van Kralingen, like us, uses Hohfeld's fundamental legal concepts as a foundation. However he chose to change the names of these concepts and dropped Hohfeld's focus on legal relations. His approach mixes up the description of *the rules of the game* and those of *playing the game*, although the latter does not include some important aspects of social interaction, e.g. agency needed to reason about the impact of norms on society. In our opinion this weakens the usefulness of his frame-based conceptual models, and resulted in an approach that is less attractive for legal experts.

In the nineties knowledge engineers focused on abstract legal ontologies and different of these legal ontology frameworks were developed [16]. The main focus of the research on such abstract formal conceptual models was on their computational abilities. How to actually make concrete conceptualizations, or legal domain ontologies from a jurisprudential, or legal perspective was listed as future work [16]. The method presented in this paper aimes to fill this gap.

3 THE CALCULEMUS APPROACH

The CALCULEMUS approach is a normative system in three layers: sources of norms in natural language (1), the formal interpretation of norms in a 'Institutional Reality' (2), and the use of a formal interpretation of norms in 'Social Reality' (3) (see figure 1).

This model is an extended version of the three layers of reality model presented in [13] and was based upon the work of Searle [9]:

1. Sources of Norms

This layer describes the components, structure and referential mechanisms that allow us to refer to the natural language sources describing the norms we want to 'translate' into formal computational models.

2. 'Institutional Reality'

This layer describes the interpretation of the sources of norms in the previous layer, using: states representing situations; legal positions; and acts regulated by norms.

3. 'Social Reality'

The 'Social Reality' layer describes agents, agent-roles, collaboration of agents, coordination, message passing, and other behavioral aspects of agents. This layer is used to describe and simulate behavior in societies regulated by norms. These norms can be used, e.g., to test (non-) compliance scenarios, and to predict effectiveness.

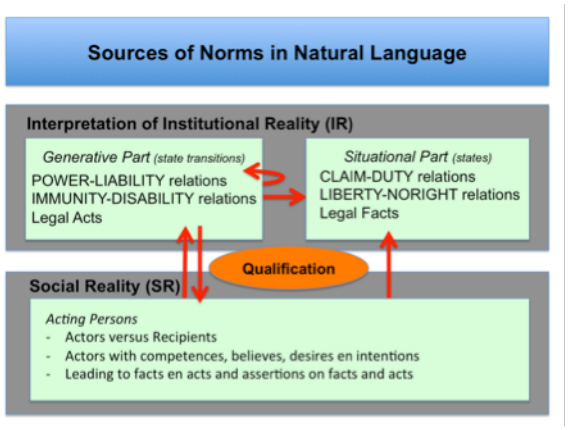


Figure 1. Three layers of reality

The second layer, 'Institutional Reality', is constructed to enable computational law. Concepts, or 'institutional facts' are derived from sources of norms, and are put in an explicit structure. Van Engers and Van Doesburg have introduced 'Institutional Reality' based on Hohfeld's fundamental legal concepts [14].

The third layer is the 'Social Reality' that can contain any brute or social fact. To qualify a social fact as a 'institutional fact' a qualified official is needed. This can be the administrator deciding on an application or objection, or it can be a judge ruling on an appeal.

The CALCULEMUS approach results in 'institutional facts' that can be used by our DSL-based reasoner build in DSL FLINT, to calculate normative positions.

'Social Reality' is modeled using agent-role modeling, see for example [10]. This paper focuses on the second layer the formal interpretation of norms, using the newly developed Formal Language for the Interpretation of Normative Theories: FLINT.

3.1 'Institutional Reality' for normative systems

'Institutional Reality' is build out of normative relations. A normative relation (NR) is an extension of the concept of a 'legal relation', as defined by Hohfeld [6]. Legal relations are based on legal sources in natural language. A normative relation can have any normative source. The elements of a 'Institutional Reality' for normative systems are described individually below.

Facts and Acts

'Institutional facts' (iFACTs) in FLINT are facts that can be traced back to a normative source. Facts in 'Social Reality' can be qualified as iFACTs.

'Institutional acts' (iACTs) are acts that can be traced back to a normative source. Acts in 'Social Reality' can be qualified as iACTs.

'Normative Relations'

Situational Normative Relations ('situational NRs') exist in two types: 'Claimright-Duty relations' and 'Liberty-Noright relations' They exist of the following elements: the 'holder' of a 'Claimright' or 'Liberty'; the holder of a 'Duty' or 'Noright'; the 'Object of the normative relation'; and the 'Duty' or 'Noright' itself. For every element of the 'situational NR' references to normative sources are registered.

'Situational NRs' are created and terminated by 'generative Normative Relations'.

Generative Normative Relations ('generative NR') also exist in two types: 'Power-Liability relations' and 'Immunity-disability relations'. They exist of the following elements: an 'Actor' (a person); a 'Recipient' (another person); an 'institutional act (iACT); the 'Object of the normative relation'; a 'precondition' and a 'postcondition'. The references to normative sources of the elements of a 'generative NR' are also registered.

The 'precondition' of a 'generative NR' is a iFACT or a set of iFACTs and 'situational NRs' combined using Boolean connections.

The 'postcondition' is a set of iFACTs and 'situational NRs' that are created and/or terminated by a 'generative NR'. The 'postcondition' can only be reached when the 'precondition' is met and an act in 'Social Reality' is qualified as the iACT belonging to the 'generative NR' that the 'postcondition' is a part of.

The 'postcondition' describes the transition of the initial state that fulfills the 'precondition' to an end state.

3.2 An example

The case study in this paper is on the admission of international students to the Netherlands. The case study is described in more detail in [14]. In this paper an example of a 'generative NR' relevant for the case study is used to present the Domain Specific Language 'FLINT'. The sources of law used in this example are English translations of the original Dutch text published on overheid.wetten.nl by the Dutch Formal Publications Office.

The 'generative NR' with code NR.AA.16.1.b is a Power-Liability relation based on article 16, paragraph 1, point b, AA. This article describes the Power of Our Minister of Justice to reject an application for a residence permit if the alien does not possess a valid border crossing document. The alien is the 'Recipient' in this NR. The 'precondition' of NR.AA.16.1.b is the existence of a application (iFACT ApplicationExists) and the existence of the iFACT that the alien does not possess a valid border crossing document (iFACT NoValidBorderCossingDoc).

The 'precondition' also exist of the absence of three exceptions for the Power to reject an application on the ground that the alien does not possess a valid border-crossing document. These exceptions are: the alien proofs that he can not (any longer) be put in possession of a valid border-crossing document due to the government of his country (1), the alien is citizen of Somalia and the Netherlands do not recognize the Somalian authorities and Somalian documents (2), and the alien is a child born in this country born who apply for stay with their parents, provided it meets the conditions for residence with its parents (3).

The first exception is described in article 3.72 of the Aliens Decree (AD. The second and third exception in chapter B1/4.2, sentence 4 of the Aliens Act Implementation Guidelines (AAIG).

4 FLINT: A Domain-Specific Language for Specifying Norms

4.1 Introduction

Domain-specific languages (DSLs) are software languages tailored toward a particular problem domain. Well-designed DSLs provide custom notation, which is closer to the concepts and vocabulary of domain experts. This improves productivity (shorter programs), quality assurance (better error message through domain-specific checks), and stakeholder communication (programs are expressed in terms of the problem domain). DSLs have been successfully applied in areas such as financial services, hardware description, and web development (for related work on DSLs, see [8]).

Although DSLs provide a bridge between a problem domain and its technical realization in software, DSL development requires both language engineering expertise and domain knowledge. Recent developments in the area of language workbenches provide integrated tool support for significantly lowering the complexity of this task [5]. Language workbenches take traditional compiler tools to the next level, by also providing support for defining editor services, such as syntax highlighting, outline views,

cross-referencing, static analyses, and documentation generation. Although often overlooked, user-level tool support is essential for adopting formal languages in non-technical domains.

In this section we present FLINT, a DSL for describing and executing norms. The current prototype of FLINT is designed as a textual language using the meta programming system Rascal [7].

4.2 FLINT

FLINT is a domain-specific formal language (DSL) that is targeted towards describing our models of 'Institutional Reality'. We will illustrate FLINT by using an example which formalizes the "reject" relation introduced in 3.3.2, see figure 2 for an example of the source code.

```
iFact ApplicationExists
  source: Article 14, paragraph 1, point a, AA
  {The application to grant a temporary residence permit}

iFact NoValidBorderCrossingDoc
  source: Article 16, paragraph 1, point b, AA
  {The alien does not possess a valid border-crossing document.}

iFact CannotHaveBorderCrossingDoc
  source: Article 3.72, AD
  {The alien proves that he can not (any longer) be put in
   possession of a valid border-crossing document due to
   the government of his country.}

iFact CitizenOfSomalia
  source: B1/4.2 sentence 4 AAIG
  {The alien is citizen of Somalia.}

iFact ChildrenBornInNL
  source: B1/4.2 sentence 4 AAIG
  {Children born in this country born who apply for stay with their parents,
   provided they meet the conditions.}

iFact RejectedBecauseNoValidBorderCrossingDoc
  source: Article 16, paragraph 1, introduction and point b, AA
  {The application to grant a temporary residence permit is rejected because the
   alien does not possess a valid border-crossing document.}

relation NR.AA.16.1.b: [Our Minister] has the power towards [the alien]
  to [reject] [the application to grant a temporary residence permit]
  source: Article 16, paragraph 1, introduction and under point b, AA
when
  ApplicationExists AND NoValidBorderCrossingDoc
  AND NOT (CannotHaveBorderCrossingDoc OR CitizenOfSomalia OR ChildrenBornInNL)
action:
  + RejectedBecauseNoValidBorderCrossingDoc
  {An application to grant a temporary residence permit as referred to in Article
   14 may be rejected if: the alien does not possess a valid border-crossing docu
```

Figure 2. Source text FLINT

The first six declarations capture the 'institutional facts' that may hold or not. Each iFACT has an intuitive name (e.g., "NoValidBorderCrossingDoc"), a reference to the legal source, and the relevant fragment of the actual text of the law. Additional meta-data, such as Juriconnect identifiers that serve as references to sources of law, are normally included, but for presentation purposes have been omitted from this example.

The final declaration describes the generative POWER relation between the 'Actor' "Our Minister" and the 'Recipient' "the alien". In this example case, it describes the power to reject an application for a temporary residence permit, on the ground of not possessing a valid border-crossing document. The 'precondition' encodes the requirement of this document (after the "when" keyword), which also describes exceptions (e.g., being a citizen of Somalia). Whenever the 'precondition' holds, the 'Actor' ("Our Minister") can enact the relation, which causes the action to be executed. In this, the action consists of adding the 'institutional fact' "RejectedBecauseNoValidBorderCrossingDoc". Enforcing a generative relation represents a transition into a new (institutional) world, where additional facts are added or existing facts are withdrawn, comparable to belief revision that is a well-known mechanism in AI.

In addition to checking for iFACTs in the 'precondition', and adding or deleting iFACTs in the 'postcondition', a generative rule can also query and create or withdraw situational Normative Relations or generative Normative Relations.

4.3 Benefits of FLINT

FLINT is accompanied with an integrated development environment (IDE), which provides editor services such as automatic syntax checking, syntax highlighting, jump-to-definition (e.g., clicking on an iFACT in a 'precondition' or 'postcondition' moves the cursor to its definition), error marking, content completion, and hover documentation. Currently, the IDE displays errors when a references iFACT or relation is not defined. In the future, we will extend this with automated cross-referencing with legal texts, more advanced consistency checks (e.g., decidability, reachability, etc.), and refactoring operations (e.g., consistently renaming an iFACT).

We have automatically imported an initial set of legal relations and iFACTs from Excel sheets, which immediately uncovered a number of mistakes due to typos or missing iFACT declarations. Automated tooling for engineering such legal specifications is thus useful, even if the current analysis support is still quite primitive.

FLINT specifications can be used for simulating cases. This involves defining an initial world by listing the iFACT and situational relations that hold. Given this initial world, some of the generative relations are enabled, because the 'precondition's are true. In the simulation, the enabled relations can be fired, to obtain a new state of the world, in which a (possibly different) set of relations is enabled.

Though the study case only includes the interpretation of one 'Normative Relation', it does show the approach to interpret these sources.

The case can be extended in two ways:

1. By collecting and interpreting normative sources of other actions than rejecting an application because the alien does not possess a valid border-crossing document: e.g. assessing other grounds for rejection, or assessing the acts of granting or disregarding an application.
2. By collecting and interpreting normative sources of normative statements that further specify the iFACTs used in the 'precondition': e.g. normative rules on establishing the iFACT that the alien does not possess a valid border-crossing document, on the definition of a border-crossing document, and on the iFACTs that determine the validity of a border-crossing document.

5 CONCLUSION AND FUTURE WORK

In this paper we made the case for formalization of normative positions and relations, separated from the formalization of agent behavior.

One of the reasons for separating the *rules of the game* from describing the *playing of the game* is that the latter is much more difficult and requires us to really understand the full complexity of intelligently operating agents in a complex adaptive systems context. While formalizing the rules of the game is a relatively easier task and the results thereof are already showed to be beneficial for practice, research on norm-guided social behavior in complex adaptive systems is still ongoing [4].

In this paper we presented the main ideas behind our CALCULEMUS method. We presented the semi-formal model of 'Institutional Reality' that is an interpretation of the sources, and we showed the formal model expressed in a DSL named FLINT (Formal Language for the Interpretation of Normative Theories).

We accept that there are still many open issues particularly in modeling 'Social Reality'. We're grateful to being able to work in spirit of the great philosopher Leibniz that initiated the idea. Calculemus!

ACKNOWLEDGEMENTS

We thank IND, DCTA, NWO, and the participants of the ICT with Industry workshop 2015 for their support and contributions.

REFERENCES

[1] T. Bench-Capon and F.P. Coenen, 'Isomorphism and legal Knowledge Based Systems', *Artificial Intelligence and Law*, 1-1, 65-86 (1991)

[2] T. Bench-Capon and T.F. Gordon, 'Isomorphism and Argumentation', in P. Casanovas (Ed.), *Proceedings of the 12th international conference on artificial intelligence and law*, 11-20, New York, NY, USA: ACM (2009).

[3] CEN MetaLex. *Open XML Interchange Format for Legal and Legislative Resources*. [Online]. Available from: http://www.metalex.eu.

[4] A. Deljoo, L.H.M. Gommans, T.M., Engers, and C.T.A.M. de Laat, 'An Agent-Based Framework for Multi-Domain service networks: Eduroam case study', in *8th International Conference ICAART 2016,* 275-280*,* Springer (2016).

[5] S. Erdweg, et al., 'Evaluating and comparing language workbenches: Existing results and benchmarks for the future', *Computer Languages, Systems & Structures*, 44, Part A, 24-47, (2015).

[6] W.N. Hohfeld and W.W. Cook, *Fundamental legal conceptions as applied in judicial reasoning, and other legal essays*, New Haven: Yale University Press, 1919.

[7] P. Klint, T. van der Storm, and J. Vinju, 'EASY meta-programming with rascal', in J.M. Fernandes, et al. (eds.), *GTTSE III. LNCS*, 6491, 222–289, Springer, Heidelberg, (2011).

[8] M. Mernik, J. Heering, A.M. Sloane, 'When and how to develop domain-specific languages', *ACM Computing Surveys (CSUR)*, 37-4, 316-344 (December 2005)

[9] J.R. Searle, *The construction of social reality*, Penguin Books, London, 1996.

[10] G. Sileno, A. Boer, and T. M. van Engers, 'Commitments, expectations, affordances and susceptibilities: Towards positional agent programming', *PRIMA 2015: Principles and Practice of Multi-Agent Systems Lecture Notes in Computer Science*, 687-696, (2015.)

[11] R. van Doesburg et al., 'Towards a Method for a Formal Analysis of Law, Study Case Report ICT with Industry workshop 2015', NWO (2016). [Online]. Available from: http://www.nwo.nl/over-nwo/organisatie/nwo-onderdelen/ew/bijeenkomsten/ict+with+ industry+workshop/proceedings

[12] T.M. van Engers, 'Legal Engineering: A Structural Approach to Improving Legal Quality', in A. Macintosh, R. Ellis, R., and T. Allen (eds.), *Applications and Innovations in Intelligent Systems XIII, proceedings of AI-2005*, 3-10, Springer, (2005).

[13] T. M. van Engers, A. Boer, 'Public Agility and Change in a Network Environment', in Judith Schossboeck, Noella Edelmann and Peter Parycek (Eds.,), *JeDEM 3(1),* 99-117, (2011).

[14] T.M. van Engers and R. van Doesburg, 'Modeling the Interpretation of Sources of Norms', in *proceedings of eKNOW2016*, IARIA XPS Press, 41-50, (2016).

[15] T.M. van Engers and E. Glassée 2001. 'Facilitating the Legislation Process Using a Shared Conceptual Model' *IEEE Intelligent Systems*, 16(1), 50-58, IEEE, (2001).

[16] R.W. van Kralingen, Frame-based conceptual models of statute law, Kluwer, (1995)

[17] P.R.S. Visser and T.J.M. Bench-Capon, 'A Comparison of Four Ontologies for the Design of Legal Knowledge Systems', *Artificial Intelligence and Law*, 6-1, 27-57, (1998).

On the Concept of Relevance in Legal Information Retrieval

Marc van Opijnen[1] and **Cristiana Santos**[2]

Abstract. The concept of 'relevance' is crucial to legal information retrieval, but because of its intuitive understanding it goes undefined too easily and unexplored too often. We discuss a conceptual framework on relevance within legal information retrieval, based on a typology of five relevance dimensions used within general information retrieval science, but tailored to the specific features of legal information. We come forward with several suggestions to improve the design and performance of legal information retrieval systems.

1 INTRODUCTION

Legal Information Retrieval (LIR) has always been a research topic within Artificial Intelligence & Law ('AI & Law'): in 'A History of AI & Law in 50 papers' [1] seven of those 50 papers have a relation to LIR. For the legal user though much research seems to be only remotely relevant for solving their daily problems in information seeking. The underrepresentation of legal practitioners within the AI & Law community might offer an explanation: *"A lawyer has always the huge text body and his degree of mastery of a special topic in mind. For a computer scientist, a high-level formalisation with many ways of using and reformulating it is the aim."*[3] Not surprisingly, LIR has been approached within AI & Law primarily with a focus on conceptualization of legal information, while for daily legal work that might not always be the most effective approach.

Meanwhile, due to the advancements of the information era and the Open Data movement the number of legal documents published online is growing exponentially, but accessibility and searchability have not kept pace with this growth rate. Poorly written or relatively unimportant court decisions are available at the click of the mouse, exposing the comforting myth that all results with the same juristic status are equal. An overload of information (particularly if of low-quality) carries the risk of undermining knowledge acquisition possibilities and even access to justice.

Apart from the problems with the quantities, also the qualitative complexities of legal search cannot easily be underestimated. Legal work is an intertwined combination of research, drafting, negotiation and argumentation. To limit the role of LIR within daily legal practice to just finding the court decisions relevant to the case at hand underestimates the complexities of the law and legal research. Any legal information retrieval system built without sufficient knowledge, not just of the actual legal information needs but also of the 'juristic mind', is apt to fail.

To aid researchers and system designers in designing or developing LIR applications it might be an interesting exercise to approach LIR more explicitly as a subtype of Information Retrieval (IR) instead of (merely) a topic within AI & Law. Since 'relevance' is the basic notion in IR, it could be a useful starting point for analysing the specificities of LIR. In this paper we develop a conceptual framework and come forward with suggestions for improvements in LIR.

In section 2 we define 'Legal Information Retrieval' by, on the one hand, distinguishing it from Legal Expert Systems and, on the other hand, describing the characteristics that justify its classification as a specific subtype of IR. In section 3 we discuss the concept of relevance in LIR, guided by a topology of five different 'dimensions' of relevance. In section 4 we will draw some conclusions and make suggestions for future work.

2 LEGAL INFORMATION RETRIEVAL

2.1 Inference Versus Querying

In a variety of ways information technology is working its way into the legal domain and even endangering the livelihood of its inhabitants.[2] Out of all these different systems we highlight two types of information systems: Legal Expert Systems (LES) and Legal Information Retrieval (LIR), on the hand with a view to articulate the particularities of LIR systems and on the other hand to underline the need – at least for many years to come – of LIR for the legal profession. The main aspects of LES and LIR are listed in table 1.

In research interesting cross-fertilisation experiments started a long way back [3] and many of the recent developments within the legal semantic web (as summarized in e.g. [4]) are also of importance for LIR, but it is highly unlikely that the two types will completely merge. LIR starts where LES isn't able to provide an answer. And notwithstanding the improvements AI & Law brings to LES, there will always be questions left and relevant documents to be discovered, since the lack of any final scheme is inherent to the legal domain.

[1] Publications Office of the Netherlands (UBR|KOOP), email: marc.opijnen@koop.overheid.nl.
[2] University of Barcelona (IDT-UAB) and University of Luxembourg, email: cristiana.teixeirasantos@gmail.com
[3] E. Schweighofer in [1, par. 2.4]

Table 1. A comparison between Legal Expert Systems (LES) and Legal Information Retrieval (LIR).

Aspect	LES	LIR
Goal	Establish a legal position on specific case	Provide relevant legal information
Input	Facts	Request
Content	Legal rules encoding the domain expertise	Documents
Method	Inference	Querying
Output	Decision, advice, forecast.	Set of documents
Preferred use	Answering 'happy flow' questions within a specific and limited domain	Finding information objects in huge repositories
Advantage	Can provide straightforward answers	Unlimited content, input and output
Drawback	What has not been modelled, cannot be answered	User always has to read, interpret and decide for himself
Basic notion	Uncertainty	Relevance

2.2 Characteristics of Legal Information

A variety of specific features justify – and compel – the positioning of Legal Information Retrieval as a specific subtype of Information Retrieval [5]. On describing these features we will briefly elucidate some shortcomings of general IR in meeting the needs arising from the legal domain.

1. *Volume.* Although in the age of 'big data' the longstanding impressive volumes of legal materials have been surpassed by e.g. telecommunications and social media data, viewed upon from an information retrieval perspective the volume of legal materials is still impressive. This holds true for public repositories (like case law databases) as well as for private repositories (e.g. case files within law firms or courts).
2. *Document size*. Compared to other domains, legal documents tend to be quite long. Although metadata and summaries are often added, access to (and searchability of) the full documents is of paramount importance.
3. *Structure*. Legal documents have very specific (internal) structures, which often also are of substantive relevance. Although standards for structuring legal documents are emerging [6], many legal documents do not have any (computer readable) structure at all.
4. *Heterogeneity of document types*. In the legal sphere a variety of document types exist which are hardly seen in other domains. Apart from the obvious legislation and court decisions, one can think of Parliamentary documents, contracts, loose-leaf commentaries, case-law notes a.s.o.
5. *Self-contained documents.* Contrary to many other domains, documents in the legal domain are not just 'about' the domain, they actually contain the domain itself and hence they have specific authority, depending on the type of document. A statute is not merely a description of what the law is, it constitutes the law itself [5]. Notwithstanding the notion that in a bibliographical sense a document is only a manifestation of an abstract work [7], for information retrieval purposes the object to be retrieved embodies the object itself.
6. *Legal hierarchy*. The legal domain itself defines a hierarchical organization with regard to the type of documents and its authority. Formal hierarchies depend on the specific jurisdiction or domain, and factual hierarchies often also depend on interpretation, e.g. the general rule *lex specialis derogat legi generalis* requires a decision on its applicability in a specific situation.
7. *Temporal aspects*. Within the incessant flow of legislative processes, legislative texts and amendments follow one another and may overlap. Recurrent challenges stem from tracing the history of a specific legal document by searching the temporal axis of its force and efficacy [8] and by retrieving the applicable law in respect to the timeframes covered by the events subject to regulation [9].
8. *Importance of citations*. In most other scientific domains citation indexes exist for academic papers. In the legal domain, citations are a more integral part of text and argumentation: *"Legal communication has two principal components: words and citations"* [10, p. 1453]. Citations can be internal (cross-references), linking one normative provision to another normative provision in the same document or normative provisions to recitals [11]. Citations can also be external, linking e.g. a court decision to a normative provision, a normative document to another normative document, or an academic work to a Parliamentary report. Citations can be explicit or implicit and they can express a whole variety of different relationships: they can be instrumental (or 'formal') – e.g. a court of appeal referring to the appealed first instance decision – or of a purely substantive nature, but having distinct intensions.
9. *Legal terminology*. Legal terminology has a rich and very specific vocabulary, characterized by ambiguity, polysemy and definitions that are very precise and vague at the same time.
10. *Audience.* Legal information is queried by a wide variety of audiences. Laymen with different levels of legal knowledge and jurists with completely different professions (e.g. scholars, judges, lawyers, notaries or legal aid workers) have completely different information retrieval needs.
11. *Personal data*. Many legal documents contain personal data. Apart from the consequences for the publication of e.g. court decisions, it also weighs on LIR, since the juristic memory is often built on names of persons and places.
12. *Multilingualism and multi-jurisdictionality*. In many (scientific) domains English is the pivotal language, and in the legal domain the same goes for common law jurisdictions. Civil law jurisdictions though have a variety of languages; language and jurisdiction have such a strong relationship that translated documents can only be a derivative of the original. As a result European or international information retrieval poses its own problems.

Strongly related to these specific characteristics of legal information, 'legal search' differs substantially from 'non legal search' [12, 13], e.g. with regard to history tracking.

3 RELEVANCE WITHIN LEGAL SEARCH

3.1 Nature of Relevance in LIR

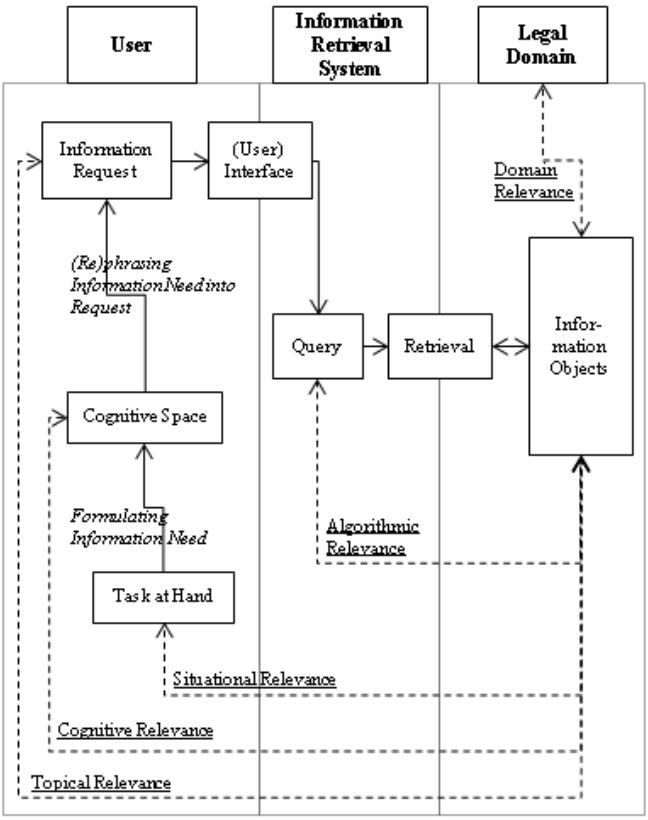
The science of Information Retrieval is basically about 'Relevance': how to retrieve the most relevant documents from – in principle – an unlimited set? Before any methodology or system for retrieval can be developed or discussed, the concept of 'relevance' has to be examined. This seems to be a trivial undertaking since this concept has a tendency to be immediately understood by everybody. A thorough understanding though is of the utmost importance for the effectiveness of LIR systems, and hence it needs continuous consideration. The foundations of a conceptual framework can be adopted from general IR science. Saracevic defined 'relevance' as: *'pertaining to the matter at hand'* [14], or, more extended: *'As a cognitive notion relevance involves an interactive, dynamic establishment of a relation by inference, with intentions toward a context.'* From this definition it follows that relevance has a contextual dependency since it is measured in comparison to the 'matter at hand'. From this definition it also follows that relevance is a comparative concept or a measure (irrelevant, weakly relevant, very relevant), which by using a specific threshold can be turned into a binary value and hence a property (relevant or not). Because of its dynamic establishment relevance may change over time and it involves some kind of selection.[15]

3.2 Dimensions of Relevance in LIR

To understand the concept of relevance it is important to disambiguate the various 'manifestations' as understood by Saracevic [15], although we prefer to use Cosijn's label 'dimensions' [16]. We discuss them here in brief, and elaborate them in the following sub-sections. The role of these dimensions in the interplay between user, information retrieval system and legal domain is depicted in figure 1.[4]

1. *Algorithmic or system relevance.* A computational relationship between a query and information objects, based on matching or a similarity between them. Traditionally models have been described within the context of full-text search, e.g. being Boolean, probabilistic, vector-space a.s.o. Natural language processing is also perceived to be within algorithmic relevance, although in our view it covers also those processes which do not take place during the actual querying, but are intended to improve algorithmic relevance at a later stage. Examples are pre-processing of documents, automatic classification a.s.o. Unlike all other relevance dimensions that can be observed and assessed without a computer, algorithmic relevance cannot: it is system-dependent.
2. *Topical relevance.* The relationship between the 'topic' (concept, subject) of a query and the information objects retrieved about that topic. A topicality relation is assumed to be an objective property, independent of any particular user. 'Aboutness' is the traditional distinctive criterion. The topics of the information objects might be hand-coded or computed, e.g. by classification algorithms. The self-containment feature of legal information adds 'isness' – as a sibling of 'aboutness' – to topical relevance in LIR.
3. *Cognitive relevance or pertinence.* Concerns the relation between the information needs of a user and the information objects. Unlike algorithmic and topical relevance, cognitive relevance is user-dependent, with criteria like informativeness, preferences, correspondence and novelty as measuring elements.

Figure 1. Interplay between user, information retrieval system and legal domain.



4. *Situational relevance or utility*. Defined as the relationship between the problem or task of the user and the information objects in the system. Also this dimension of relevance is dependent on the specific user, but unlike the cognitive relevance it does not focus on the request as formulated, but on the underlying motivation for starting the information retrieval process. Inferred criteria for situational relevance are the usefulness for decision-making, appropriateness in problem solving and reduction of uncertainty.
5. *Domain relevance.* As the fifth dimension Saracevic [14] used 'Motivational or affective relevance', but in a critical assessment Cosijn and Ingwersen [16] replaced this dimension by 'socio-cognitive relevance', which *"[I]s measured in terms of the relation between the situation, work task or problem at hand in a given socio-cultural context and the information objects, as perceived by one or several cognitive agents."* Given the specific features of legal information as well as for reasons of modelling, we define this dimension as the relevance of information objects within the legal domain itself

[4] Inspired by [17].

(and hence not to 'work task or problem at hand'). For convenience we label it 'domain relevance'.

Already here it should be observed that relevance dimensions easily overlap and intermingle: *"The effectiveness of IR depends on the effectiveness of the interplay and adaptation of various relevance manifestations, organized in a system of relevancies."*[14, p. 216] In the design of IR systems it is hence of the utmost importance to distinguish between its various dimensions and to pay specific attention to each of them, in the user interface, the retrieval engine and the document collection. It will definitely improve the user's perception of the system's performance on retrieving the most relevant information. This perception – or 'criterion for success' – depends on the relevance dimension(s) invoked. These criteria are, together with the nature of the respective dimensions, summarized in table 2.

Table 2. Dimensions of Relevance Compared

Dimension	Describes a relation between	Criterion for 'success'
Algorithmic relevance	Query and information objects	Comparative effectiveness in inferring relevance
Topical relevance	Topic or bibliographical object expressed in the request and topic or bibliographical object covered by information objects	Isness / aboutness
Cognitive relevance	Information needs of the user and information objects	Cognitive correspondence, novelty, information quality, authoritativeness, informativeness
Situational relevance	Situation / task at hand	Usefulness in decision-making and problem-solving
Domain relevance	Opinion of the legal crowd and information objects	Legal importance / authority

In the following subsections we will elaborate the five relevance dimensions in LIR and discuss how these dimensions may help to classify past and current spectrum of approaches and how it might help bridging the conceptual gap between lawyers and informaticians.

3.2.1 Algorithmic Relevance

Algorithmic relevance concerns the computational core of information retrieval. As expressed in figure 1 it is the relation between the information objects and the query; this 'query' is to be understood as the computer processable translation of the request as entered in the user interface or any other intermediary component. Algorithmic relevance is about the capability of the engine to retrieve a given set of information objects (the 'gold standard') that should be retrieved with a given query (measured in 'recall') with a minimum of false positives (measured in 'precision').

From our conceptual perspective the type of query as well as the type of retrieval framework is not relevant, but given the legal information features of volume, document size and lack of structure, textual search has for long had the focus. In the early days Boolean search was the core of any legal retrieval system, and it is still an indispensable element in most LIR systems today. In a Boolean system both the user request and the documents are regarded as a set of terms, and the system will return documents where the terms in the query are present. Boolean searches often result in the retrieval of a large number of documents. In addition, they provide little or no assistance to the user in formulating or refining a query and they lack domain expertise that could improve the search outcome. Relevance performance was improved by using models as the *vector space model* [18] and *TF-IDF* (term frequency – inverse document frequency). Nevertheless, recall is often below acceptable levels because the design of full-text retrieval systems: *"(I)s based on the assumption that it is a simple matter for users to foresee the exact words and phrases that will be used in the documents they will find useful, and only in those documents."* [19]. Ambiguity, synonymy and complexity of legal expressions contribute substantially to this problem.[20] Natural language processing (NLP) is gaining popularity as an addition to or alternative to pure text-based search.[21]

Apart from text-based search also other types of calculations can be considered within 'algorithmic relevance', like the use of ontologies as higher level knowledge models [4, 22] as well as network statistics, especially when used for citation analysis [23, 24].

3.2.2 Topical Relevance

Topical relevance is about the relevancy relation between the topic as formulated in the user request and the topic of the information objects. But before we can discuss this 'traditional' topicality within LIR, attention should be drawn to an often overlooked feature of legal information that is of crucial importance for topical relevance: its self-containment. A classic car database contains documents *about* classic cars, not the cars themselves, while a legislation database does contain the legislative texts themselves. Because the same repository might also contain other acts or court decisions citing it or scholarly writings discussing it, we add 'isness' to 'aboutness' as a separate concept within topical relevance. We will discuss both concepts below.

Isness

In general two types of searching can be distinguished: searching the known and searching the unknown. Searching the known in LIR concerns 'isness': finding a specific law, court case, Parliamentary document or scholarly article, generally by keying in some kind of identifier (e.g. a title or a code). Although this might look like a problem of 'data retrieval' instead of 'information retrieval' [25, par. 1.1.1] and hence a no-brainer [26], in most legal information systems it is still a real brainteaser. A first reason for this is that 'isness' is too easily confused with 'an exact match' or a specific document while, on the contrary, a whole set of different documents can be correctly retrieved by an isness request: an initial act as published in an Official Journal, as well as a series of consolidated versions, all in different language expressions and/or in different formats. A second reason is that lawyers often refer to the work level, [7] while the search engine is not clever enough to relate the work level to the actual information objects. A third reason is the improper or incomplete pre-parsing of the user

request, resulting in interpreting the request as text query instead of understanding it as an identifier for a (series of) information objects.

This can e.g. be illustrated in EUR-Lex: using the quick search field for searching by document number (*'Regulation (EEC) No 1408/71'*), often used formatting variants (*'Reg. 1408/71'*) or aliases (*'Services directive', 'Dublin Regulation'*) does not render the documents which are identified by these labels on top of the result list, and when using the advanced search – where one has to split the document number into a 'year' part and a 'number' part – the non-specialist user is probably puzzled where to put which digits for 'Directive 96/95/EC', 'Regulation 98/2014' or 'Regulation 2015/2016'.[5]

To improve topical relevance it is important to understand that a user of legal information retrieval systems generally prefers – if possible – isness over aboutness. To achieve such improvement isness should always rank higher in a result list than aboutness, or even better: be labelled as such. Secondly, the capabilities of the system should be improved as to recognize requests that refer to isness, including all the many ways in which isness can be expressed, such as: different types of identifiers for the same thing, many different formatting styles even for one type of identifier, the use of 'global aliases' like 'Bolkestein Directive' or 'General Data Protection Regulation'. Reference parsers have been developed for detecting citations in the documents themselves [27] (see below, section 3.2.5), but can also be used for parsing user requests.

Aboutness

While 'isness' relates to searching the known, 'aboutness' relates to searching the unknown: one is not searching for a specific document (or work), but for information or knowledge about something.

Using free text search and mapping the searched terms to the terms indexed from the information objects renders poor results since legal concepts can be expressed in a variety of ways, while completely different concepts can textually be quite similar.

Adding head notes and keywords from taxonomies or thesauri has been a long tradition within the legal information industry. Although aboutness is assumed to be an objective property, independent of any particular user, manual indexing is inherently subjective, and even the same indexer may sort the same document under different terms depending on which context the document is presented in [28]. *"Manual indexing is only as good as the ability of the indexer to anticipate questions to which the indexed document might be found relevant. It is limited by the quality of its thesaurus. It is necessarily precoordinated and is thus also limited in its depth. Finally, like any human enterprise, it is not always done as well as it might be."*[20, p. 14] Semi-automated classification using ontologies [29] is gaining popularity, but automatic classification turns out not to perform better than human indexing.[30] For huge public databases manual tagging is hardly an option. And a general drawback of such systems is the mandatory use in the user interface of the classification scheme. This forces the user to limit or to reformulate his request to align it with the available classification system. A problem that can only be solved by the time-consuming and tedious task of *"Using a combination of automated and manual techniques, [constructing] a list of concepts and variations for expressing a concept."*[6] This requires in-depth legal knowledge, analysis of search engine log files and continuous maintenance.

Search in multilingual legal repositories – e.g. in the ECLI Search Engine on the European e-Justice portal[7] – poses specific problems: the terms used in the request do not only have to be translated into the language of the information objects, but also into the specific legal terminology of the jurisdiction the information objects are about. Sufficient solutions have not yet been developed. EuroVoc[8] is a large multilingual vocabulary; although it is used for tagging in the EUR-Lex database, it too much policy-oriented and too less legal to be of practical use for LIR. Aligning legal vocabularies of different legal systems and/or languages has proven to be quite difficult [31]. Within the Legivoc project various national legal vocabularies have been mapped [32], but it needs more elaboration to be of practical use.

Meanwhile, developers of LIR systems should consider whether the investment is worth the effort: surveys have shown that classification systems are not very popular among users [33], contrary to searches by relationship [34]. Many topics in law, at least in the juristic mindset and information seeking behaviour, have a strong connection to other legal documents. Typical requests may refer to a search for (everything) about a specific paragraph of law or court decision. In such requests these information objects represent a specific legal concept, but the only reason lawyers rephrase it might be related to the fact that the search engine cannot cope with their actual request. For well-known acts and codes such aboutness information is structured in treatises or loose-leaf encyclopaedias, but they are optimized for browsing, not for search. Since such works do not cover the whole legal domain, performing searches on citations might in principle be the obvious choice.

In common law countries citators are very popular for such 'topical citation search', like LawCite.org in the public domain and Shepard's in the private domain. The latter is based on manual tagging and also contains qualifications of these relations. In continental Europe the importance of search by citation – as a type of aboutness – needs more attention from search providers. In EUR-Lex, HUDOC and various national legislative databases, relations between documents are tagged and searchable/browsable, but especially in national case law databases citation search is extremely difficult. A first reason is that judges have lousy citations habits: research showed that only 36% of cited EU acts was in conformity with the prescribed citation style, the other citations were made with a wide range of other styles [35]. Comparable problems appear when searching for case law citations, where additional complexity is added by the fact that one decision can be cited by many different identifiers.[36] Also, slashes, commas and hyphens are essential elements of legal identifiers, but are out-of-the-box interpreted by search engines as specific search instructions (like 'near' for '/' or 'not' for '-'). Manual tagging for large scale public databases is undoable, so reference parsers have to be developed [27]; as indicated in section

[5] The year is 96, 2014 and 2015 respectively. If the citation the user has at its disposal is correct and if he is knowledgeable about EU document numbering he can solve the problem, but often citations are incomplete or poorly formatted [27]. In a directive the year comes first, in a regulation the number comes first. But from 1 January 2015 onwards the year comes first in all acts <eur-lex.europa.eu/content/tools/elaw/OA0614022END.pdf>.

[6] P. Zhang, key-note speech on ICAIL 2015 Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts. <www.lrdc.pitt.edu/ashley/icail2015nlp/>.

[7] <https://e-justice.europa.eu/content_ecli_search_engine-430-en.do>.

[8] <eurovoc.europa.eu>.

3.2.1 they can be used both for recognizing the citations in the information objects as well as in the request.

3.2.3 Cognitive Relevance

Cognitive relevance concerns the extent to which an information object matches the cognitive information needs of the user. Unlike algorithmic and topical relevance, this dimension is of a subjective nature: do the retrieved documents fit to the user's state of knowledge? Is he offered the temporal version that matches his information need? Are there any metadata regarding the information objects retrieved he should be aware of?

Since this dimension is of subjective nature, the cognitive relevance performance of a LIR system depends, broadly speaking, on the continuum between on the one hand, system features to tailor the search experience to personal needs, and on the other hand, the ability of the system to explicitly or implicitly understand the information needs of each individual user. An example of the former is time travelling: jurists often need to know the delta between the temporal version T of an act and version T+1. Up until recently many legislative databases were only able to serve version T and T+1 in parallel, without actually showing the delta. By offering such functionality,[9] a user is served in his personal information needs, although the information retrieved is the same for all users.

On the other end of the continuum we find systems for personalized search, using filters to recommend information objects that are deemed relevant for a specific user at a specific stage in his information collecting process. Within such 'recommender systems' two types of filtering can be distinguished. 'Collaborative filtering' recommends documents by making use of the user's past search behaviour and/or that of a peer group. 'Content-based filtering' uses shared features of the document at hand and other documents, based on e.g. topical resemblance, having comparable metadata or closeness in a citation network. Recommender mechanisms can be used to limit the number of documents retrieved (e.g. because the systems knows user A is only interested in tax law and not in criminal law) or to increase the number of documents: by offering 'more like this' buttons or navigable citation graphs users can be supported in serendipitous information discovery.[37] Being tailored to the individual need of the user, recommender system can also be used for pro-active search: notification systems informing a user about information objects that have been added to the repository and might be of interest for him. Within legal information retrieval recommender systems have not had too much attention yet. [38]

3.2.4 Situational Relevance

While cognitive relevance is associated with search task execution, situational relevance pertains to work task execution; the relevance of documents is measured by their usefulness for the task at hand, e.g. decision-making or problem-solving.[17] It should be noted that the system is not asked to solve the problem itself – then it would be a legal expert system, not a LIR system.

Situational relevance in legal information retrieval comes close to – but should not be confused with – 'legal relevance', which usually means that information is relevant to a proposition when it affects, positively or negatively, the probability that the proposition is true [39, p. 148].[10]

The difference between 'legal relevance' and situational relevance can be understood with the help of the following definition by Jon Bing:

A legal source is relevant if:

1. *The argument of the user would have been different if the user did not have any knowledge of the source, i.e. at least one argument must be derived from the source; or*
2. *legal meta-norms require that the user considers whether the source belongs to category (1); or*
3. *the user himself deems it appropriate to consider whether the source belongs to category (1).*[41]

In this definition (1) pertains to the strict notion of 'legal relevance', while situational relevance in legal information retrieval also covers (2) and (3).

Probably because of the relative importance of case law in the United States and other common law countries, much LIR research has concentrated on finding the (most) relevant court decisions relating to a case at hand. This can be pursued using a variety of (sometimes combined) technologies, like argumentation mining [42] and natural language processing (NLP) [21].

Navigational features of LIR systems, like memorized search history, storage of relevant documents found, shared folders and customization features do not pertain to situational relevance in an IR sense, unless these data are used for collaborative or content-based recommendations that pertain to the dossier at hand.

3.2.5 Domain Relevance

We defined 'domain relevance' as the relevance of information objects within the legal domain itself. It is independent from an information system and independent from any user request. As can be understood from the previous paragraph we prefer to avoid the term 'legal relevance', but 'legal authority' or 'legal importance' are safe to use as synonyms for 'domain relevance'.

Domain relevance can be applied in LIR systems in different ways. First, it can be used to classify categories of information objects as to their legal weight: a constitution outweighs an ordinary act, which in turns is of more importance than a by-law or ministerial degree. In the same way an opinion of a supreme court can be expected to have more authority than a district court verdict, but it can be superseded by a judgment of the European Court of Human Rights.

Secondly, the concept of domain relevance can be used to classify individual information objects as to their legal authority. Separating the wheat from the chaff has for long been the territory of domain experts: since publication / storage was expensive, and adding documents itself labour-intensive, a selection was made on the input side of any paper or early digital repository. The ease with which information can now be published on the internet has shifted the issue of selection – at least partially – from the input side to the output side: 'selection' has evolved from a publisher's issue into a search issue. Case law publication in the Netherlands could serve as an illustration: the public case law database in the Netherlands[11] contains a small percentage (less than 1%) of decided cases, but in fifteen years has accumulated 370.000

[9] E.g. <wetten.overheid.nl/BWBR0006368/2016-01-01?VergelijkMet=BWBR0006368%3fg%3d2010-02-01>.

[10] Next to this 'logical' or 'probablistic' definition often also a 'practical' concept is used, meaning 'worth hearing'. [40]

[11] <uitspraken.rechtspraak.nl>.

documents. More than 75% of those are not considered important enough to be published in legal magazines.[43]

An example of domain relevance applied at the document level can be observed in the HUDOC database, containing all case law documents produced by the European Court of Human Rights. To aid the user in filtering the nearly 57.000 documents as to their legal authority, four importance levels have been introduced. Except for the highest category, containing all judgments published in the Court Reports, all documents have been tagged manually. Since this importance level is an attribute of each individual document, it can easily be used in combination with other relevance dimensions.

Since manual tagging is labour-intensive, for more massive repositories a computer-aided rating is indispensable. Given the abundant use of citations between court decisions, network analysis is an obvious methodology to assess case law authority [23, 44]. In the 'Model for Automated Rating of Case law' [24] the 'legal crowd' – the domain specialists that rate individual court decisions as to their authority by citing them or not – is extended to legal scholars, while it also uses other variables within regression analysis to predict the odds of a decision rendered today for being cited in the future. It also takes into account changing perceptions over time (see e.g. also [45]). If court decisions are well-structured and citations are made to the paragraph level, importance can be calculated for the sub-document level as well [46]. Comparable techniques can be used for the relevance classification of legislative documents [47] or for a network containing different types of sources [48].

Network analysis is supported by the use of common identifiers, like the European Legislation Identifier,[12] the European Case Law Identifier[13] [49] and possibly in the future a European Legal Doctrine Identifier (ELDI) [50] or a global standard for legal citations.[14]

4 CONCLUSIONS AND FUTURE WORK

Relevance, the basic notion of information retrieval *"Is a thoroughly human notion and as all human notions, it is somewhat messy."* [15] As upheld in this paper, 'relevance' within legal information retrieval deserves specific attention, due to rapidly growing repositories, the distinct features of legal information objects and the complicated tasks of legal professionals.

Because most LIR systems are designed by retrieval specialists without comprehensive domain knowledge, sometimes assisted by domain specialists with too little knowledge of retrieval technology, users are often disappointed by their relevance performance.

Four main conclusions can be highlighted. First of all, retrieval engineering is focussed too exclusively on algorithmic relevance, but it has been proven sufficiently that without domain specific adaptations every search engine will disappoint legal users. By unravelling the holistic concept of 'relevance' we hope to stimulate a more comprehensive debate on LIR system design. All dimensions of relevance have to be considered explicitly while designing all components of LIR systems: document pre-processing, (meta)data modelling, query building, retrieval engine and user interface. Within the user interface searching, filtering and browsing should take full advantage of the various relevance dimensions, of course in a way that fits the legal mindset and acknowledging that relevance dimensions are continually interacting in the process of information searching.

Secondly, the 'isness' concept is too often overlooked. Finding (the expressions of) a work is – and not (just) the related works – is an often-used functionality for jurists, but misunderstood by system developers.

Thirdly, domain relevance is also an underdeveloped area. While there is a tendency to publish ever more legal information, especially court decisions, without tagging it as to its juristic value, information overkill will become a serious threat to the accessibility of such databases. Performance on other relevant dimensions will suffer if the problem of domain relevance isn't tackled.

Finally, given the importance of digital information for legal professionals – lawyers easily spend up to fifteen hours per week on search, most of it in electronic resources [34] although the abandonment of paper does not always seem to be a voluntary choice [51] – the gap between LIR systems and user needs is still big. For a full understanding of their search needs just taking stock of their wishes is not going to suffice, since they are not capable of describing the features of a system that does not yet exist. To understand the juristic mindset it is of the utmost importance to follow meticulously their day-to-day retrieval quests. It will for sure reveal interesting insights that can be used to improve the relevance performance of LIR systems.

[12] Council conclusions inviting the introduction of the European Legislation Identifier (ELI), CELEX: 52012XG1026(01).

[13] Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law, CELEX: 52011XG0429(01).

[14] LegalCiteM: <www.oasis-open.org/committees/legalcitem/>.

REFERENCES

[1] T. Bench-Capon, M. Araszkiewicz, K. Ashley, et al., *A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law*, Artificial Intelligence and Law 20 (2012): 215-319.

[2] R. Susskind, *Tomorrow's Lawyers: An Introduction To Your Future*, Oxford University Press, 2013.

[3] E. L. Rissland and J. J. Daniels, *A hybrid CBR-IR approach to legal information retrieval*, *Proceedings of the 5th international conference on Artificial intelligence and law*, ACM, 1995, pp. 52-61.

[4] P. Casanovas, M. Palmirani, S. Peroni, et al., *Special Issue on the Semantic Web for the Legal Domain Guest Editors' Editorial: The Next Step*, Semantic Web Journal 7 (2016): 213-227.

[5] H. Turtle, *Text Retrieval in the Legal World*, Artificial Intelligence and Law 3 (1995): 5-54.

[6] M. Palmirani, *Legislative XML: Principles and Technical Tools*, ARACNE, Rome, 2012.

[7] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*UBCIM Publications - New Series Vol 19.

[8] M. Araszkiewicz, *Time, Trust and Normative Change. On Certain Sources of Complexity on Judicial Decision-Making*, in P. Casanovas, U. Pagallo, M. Palmirani and G. Sartor, eds., *AI*

Approaches to the Complexity of Legal Systems: AICOL 2013, Springer, 2014, pp. 100-114.

[9] M. Palmirani and R. Brighi, *Time Model for Managing the Dynamic of Normative System*, in M. Wimmer, H. Scholl, Å. Grönlund and K. Andersen, eds., *Electronic Government; Lecture Notes in Computer Science*, Springer, Heidelberg, 2006, pp. 207-218.

[10] F. R. Shapiro, *The Most-Cited Articles from The Yale Law Journal*, Yale Law Journal 100 (1991): 1449.

[11] L. Humphreys, C. Santos, L. d. Caro, et al., *Mapping Recitals to Normative Provisions in EU Legislation to Assist Legal Interpretation*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 41-49.

[12] S. Makri, *A study of lawyers' information behaviour leading to the development of two methods for evaluating electronic resources*, 2008.

[13] S. Davidson, *Way Beyond Legal Research: Understanding the Research Habits of Legal Scholars*, Law Library Journal 102 (2010): 561-579.

[14] T. Saracevic, *Relevance Reconsidered*, *Information science: Integration in perspectives. Second Conference on Conceptions of Library and Information Science*, 1996, pp. 201-218.

[15] T. Saracevic, *Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance*, Journal of the American Society for Information Science and Technology 58 (2007): 1915-1933.

[16] E. Cosijn and P. Ingwersen, *Dimensions of relevance*, Information Processing and Management 36 (2000): 533-550.

[17] E. Cosijn and T. Bothma, *Contexts of relevance for information retrieval system design*, *Proceedings of the 5th international conference on Context: Conceptions of Library and Information Sciences*, Springer-Verlag, 2154833, 2005, pp. 47-58.

[18] G. Salton, A. Wong and C. S. Yang, *A vector space model for automatic indexing*, Communications of the ACM 18 (1975): 613-620.

[19] D. C. Blair and M. E. Maron, *An evaluation of retrieval effectiveness for a full-text document-retrieval system*, Communications of the ACM 28 (1985): 289-299.

[20] D. P. Dabney, *The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval*, Law Library Journal 78 (1986): 5-40.

[21] K. T. Maxwell and B. Schafer, *Concept and Context in Legal Information Retrieval*, in E. Francesconi, G. Sartor and D. Tiscornia, eds., *Legal Knowledge and Information Systems - JURIX 2008: The Twenty-First Annual Conference*, IOS Press, Amsterdam, 2008, pp. 63-72.

[22] M. Saravanan, B. Ravindran and S. Raman, *Improving legal information retrieval using an ontological framework*, Artificial Intelligence and Law 17 (2009): 101-124.

[23] J. H. Fowler and S. Jeon, *The authority of Supreme Court precedent*, Social Networks 30 (2008): 16-30.

[24] M. van Opijnen, *A Model for Automated Rating of Case Law*, *Fourteenth International Conference on Articifial Intelligence and Law*, ACM, New York, 2013, pp. 140-149.

[25] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Pearson Education, Essex, 1999.

[26] T. Harvold, *Is searching the best way to retrieve legal documents?*, *e-Stockholm '08 Legal Conference*, 2008.

[27] M. van Opijnen, N. Verwer and J. Meijer, *Beyond the Experiment: the eXtendable Legal Link eXtractor*, *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL)*, 2015.

[28] J. Bing and T. Harvold, *Legal Decisions and Information Systems*, Universitets Forlaget, Oslo, 1977.

[29] G. Boella, L. Di Caro, L. Humphreys, et al., *Eunomos, a legal document and knowledge management system for the web*, Semantic Web Journal.

[30] S. N. Mart, *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, Law Library Journal 102 (2010): 221-249.

[31] E. Francesconi and G. Peruginelli, *Semantic Interoperability among Thesauri: A Challenge in the Multicultural Legal Domain*, in W. Abramowicz, R. Tolksdorf and K. Węcel, eds., *Business Information Systems Workshops: BIS 2010 International Workshops, Berlin, Germany, May 3-5, 2010.* , Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 280-291.

[32] H.-J. Vibert, P. Jouvelot and B. Pin, *Legivoc – connecting laws in a changing world*, Journal of Open Access to Law 1 (2013).

[33] L. F. Peoples, *The Death of the Digest and the Pitfalls of Electronic Research: What Is the Modern Legal Researcher to Do?*, Law Library Journal 97 (2005): 661-679.

[34] S. A. Lastres, *Rebooting Legal Research in a Digital Age*, 2015.

[35] M. van Opijnen, *Searching for References to Secondary EU Legislation*, in S. Tojo, ed., *Fourth International Workshop on Juris-informatics (JURISIN 2010)*, 2010.

[36] M. van Opijnen, *Canonicalizing Complex Case Law Citations*, in R. Winkels, ed., *Legal Knowledge and Information Systems - JURIX 2010: The Twenty-Third Annual Conference*, IOS Press, Amsterdam, 2010, pp. 97-106.

[37] E. Toms, *Serendipitous Information Retrieval*, *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[38] R. Winkels, A. Boer, B. Vredebregt, et al., *Towards a Legal Recommender System 27th International Conference on Legal knowledge and information systems (JURIX 2014)*, IOS Press, 2014, pp. 169-178.

[39] R. Cross and N. Wilkins, *An Outline of the Law of Evidence*, Butterworths, London, 1964.

[40] J. Woods, *Relevance in the Law: A Logical Perspective*, in D. M. Gabbay, P. Canivez, S. Rahman and A. Thiercelin, eds., *Approaches to Legal Rationality*, Springer, Dordrecht, 2010.

[41] J. Bing, *Handbook of Legal Information Retrieval*, Norwegian Research Center for Computers and Law, Oslo, 1991.

[42] R. Mochales and M.-F. Moens, *Argumentation Mining*, Artificial Intelligence and Law 19 (2011): 1-22.

[43] M. van Opijnen, *Op en in het web. Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd*, Amsterdam UvA, Den Haag, 2014.

[44] R. Winkels, J. de Ruyter and H. Kroese, *Determining Authority of Dutch Case Law*, in K. M. Atkinson, ed., *Legal Knowledge and Information Systems. JURIX 2011: The Twenty-Fourth International Conference.*, IOS Press, Amsterdam, 2011, pp. 103-112.

[45] F. Tarissan and R. Nollez-Goldbach, *Temporal Properties of Legal Decision Networks: A Case Study from the International Criminal Court*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 111-120.

[46] Y. Panagis and U. Šadl, *The Force of EU Case Law: A Multidimensional Study of Case Citations*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 71-80.

[47] P. Mazzega, D. Bourcier and R. Boulet, *The network of French legal codes*, *12th International Conference on Artificial Intelligence and Law*, 2009, pp. 236-239.

[48] M. Koniaris, I. Anagnostopoulos and Y. Vassiliou, *Network Analysis in the Legal Domain: A complex model for European Union legal sources*, arXiv:1501.05237 [cs.SI]2015).

[49] M. van Opijnen and A. Ivantchev, *Implementation of ECLI - State of Play*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 165-168.

[50] M. van Opijnen, *The European Legal Semantic Web: Completed Building Blocks and Future Work*, *European Legal Access Conference*, 2012.

[51] C. C. Kuhlthau and S. L. Tama, *Information Search Process of Lawyers: a Call for `Just For Me' Information Services*, journal of Documentation 57 (2001): 25-43.

Formalizing correct evidential reasoning with arguments, scenarios and probabilities

Bart Verheij[1]

Abstract. Artificial intelligence research on reasoning with criminal evidence in terms of arguments, hypothetical scenarios, and probabilities inspired the approach in this paper. That research showed that Bayesian Networks can be used for modeling arguments and structured hypotheses. Also well-known issues with Bayesian Network were encountered: More numbers are needed than are available, and there is a risk of misinterpretation of the graph underlying the Bayesian Network, for instance as a causal model. The formalism presented here is shown to correspond to a probabilistic interpretation, while answering these issues. The formalism is applied to key concepts in argumentative, scenario and probabilistic analyses of evidential reasoning, and is illustrated with a crime investigation example.

1 Introduction

Establishing what has happened in a crime is often not a simple task. In the literature on correct evidential reasoning, three structured analytic tools are distinguished: arguments, scenarios and probabilities [1, 8, 11]. These tools are aimed at helping organize and structure the task of evidential reasoning, thereby supporting that good conclusions are arrived at, and foreseeable mistakes are prevented.

In an *argumentative analysis*, a structured constellation of evidence, reasons and hypotheses is considered. Typically the evidence gives rise to reasons for and against the possible conclusions considered. An argumentative analysis helps the handling of such conflicts. The early twentieth century evidence scholar John Henry Wigmore is a pioneer of argumentative analyses; cf. his famous evidence charts [38, 39].

In a *scenario analysis*, different hypothetical scenarios about what has happened are considered side by side, and considered in light of the evidence. A scenario analysis helps the coherent interpretation of all evidence. Scenario analyses were the basis of legal psychology research about correct reasoning with evidence [2, 16, 37].

In a *probabilistic analysis*, it is made explicit how the probabilities of the evidence and events are related. A probabilistic analysis emphasises the various degrees of uncertainty encountered in evidential reasoning, ranging from very uncertain to very certain. Probabilistic analyses of criminal evidence go back to early forensic science in the late nineteenth century [23] and have become prominent by the statistics related to DNA profiling.

In a Netherlands-based research project,[2] artificial intelligence techniques have been used to study connections between these three tools [34]. This has resulted in the following outcomes:

- A method to manually design a Bayesian Network incorporating hypothetical scenarios and the available evidence [35];
- A case study testing the design method [35];
- A method to generate a structured explanatory text of a Bayesian Network modeled according to this method [36];
- An algorithm to extract argumentative information from a Bayesian Network modeling hypotheses and evidence [25];
- A method to incorporate argument schemes in a Bayesian Network [24].

Building on earlier work in this direction [9, 10], these results show that Bayesian Networks can be used to model arguments and structured hypotheses. Also two well-known issues encountered when using Bayesian Networks come to light:

- A Bayesian Network model typically requires many more numbers than are reasonably available;
- The graph model underlying a Bayesian Network is formally well-defined, but there is the risk of misinterpretation, for instance unwarranted causal interpretation [7] (see also [15]).

Building on the insights of the project, research has started on addressing these issues by developing an argumentation theory that connects critical arguments, coherent hypotheses and degrees of uncertainty [31, 32, 34]. The present paper expands on this work by proposing a discussion of key concepts used in argumentative, scenario and probabilistic analyses of reasoning with evidence in terms of the proposed formalism. The idea underlying this theoretical contribution is informally explained in the next section. The crime story of Alfred Hitchcock's famous film 'To Catch A Thief', featuring Cary Grant and Grace Kelly (1955) is used as an illustration.

2 General idea

The argumentation theory developed in this paper considers arguments that can be presumptive (also called ampliative), in the sense of logically going beyond their premises. Against the background of classical logic, an argument from premises P to conclusions Q goes beyond its premises when Q is not logically implied by P. Many arguments used in practice are presumptive. For instance, the prosecution may argue that a suspect was at the crime scene on the basis of a witness testimony. The fact that the witness has testified as such does not logically imply the fact that the suspect was at the crime scene. In particular, when the witness testimony is intentionally false, based on inaccurate observations or inaccurately remembered, the suspect may not have been at the crime scene at all. Denoting the witness testimony by P and the suspect being at the crime scene as Q, the

[1] Artificial Intelligence, University of Groningen, www.ai.rug.nl/~verheij

[2] See http://www.ai.rug.nl/~verheij/nwofs/.

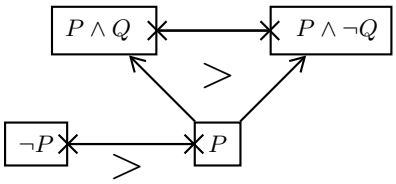


Figure 1. Some arguments

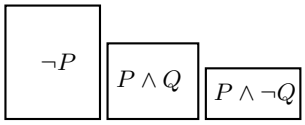


Figure 2. Some cases

argument from P to Q is presumptive since P does not logically imply Q. For presumptive arguments, it is helpful to consider the *case made by the argument*, defined as the conjunction of the premises and conclusions of the argument [29, 30]. The case made by the argument from P to Q is $P \wedge Q$, using the conjunction of classical logic. An example of a non-presumptive argument goes from $P \wedge Q$ to Q. Here Q is logically implied by $P \wedge Q$. Presumptive arguments are often defeasible [17,26], in the sense that extending the premises may lead to the retraction of conclusions.

Figure 1 shows two presumptive arguments from the same premises P: one supports the case $P \wedge Q$, the other the case $P \wedge \neg Q$. The $>$-sign indicates that one argument makes a stronger case than the other, resolving the conflict: the argument for the case $P \wedge Q$ is stronger than that for $P \wedge \neg Q$. The figure also shows two assumptions P and $\neg P$, treated as arguments from logically tautologous premises. Here the assumption $\neg P$ makes the strongest case when compared to the assumption P. Logically such assumptions can be treated as arguments from logical truth $\top$. These four arguments—two arguments implicitly from $\top$, and two from P— make three cases: $\neg P$, $P \wedge Q$ and $P \wedge \neg Q$ (the boxes in Figure 2). The sizes of the boxes suggest a preference relation.

The comparison of arguments and of cases are closely related in our approach, which can be illustrated as follows. The idea is that a case is preferred to another case if there is an argument with premises that supports the former case more strongly than the latter case. Hence, in the example in the figures, $\neg P$ is preferred to both $P \wedge Q$ and $P \wedge \neg Q$, and $P \wedge Q$ is preferred to $P \wedge \neg Q$. Conversely, given the cases and their preferences, we can compare arguments. The argument from P to Q is stronger than from P to Q' when the best case that can be made from $P \wedge Q$ is preferred to the best case that can be made from $P \wedge Q'$.

3 Formalism and properties

We use a classical logical language L with BNF specification $\varphi ::= \top \mid \bot \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \leftrightarrow \psi$, and the associated classical, deductive, monotonic consequence relation, denoted $\models$. We assume a language generated by a finite set of propositional constants.

First we define case models, formalizing the idea of cases and their preferences. The cases in a case model must be logically consistent, mutually incompatible and different; and the comparison relation must be total and transitive (hence is what is called a total preorder, commonly modeling preference relations [21]).

Definition 1 *A* case model *is a pair $(C, \geq)$ with finite $C \subseteq L$, such that the following hold, for all φ, ψ and $\chi \in C$:*

1. $\not\models \neg\varphi$;
2. *If* $\not\models \varphi \leftrightarrow \psi$, *then* $\models \neg(\varphi \wedge \psi)$;
3. *If* $\models \varphi \leftrightarrow \psi$, *then* $\varphi = \psi$;
4. $\varphi \geq \psi$ *or* $\psi \geq \varphi$;
5. *If* $\varphi \geq \psi$ *and* $\psi \geq \chi$, *then* $\varphi \geq \chi$.

The strict weak order $>$ standardly associated with a total preorder $\geq$ is defined as $\varphi > \psi$ if and only if it is not the case that $\psi \geq \varphi$ (for φ and $\psi \in C$). When $\varphi > \psi$, we say that φ is (strictly) preferred to ψ. The associated equivalence relation $\sim$ is defined as $\varphi \sim \psi$ if and only if $\varphi \geq \psi$ and $\psi \geq \varphi$.

Example. Figure 2 shows a case model with cases $\neg P$, $P \wedge Q$ and $P \wedge \neg Q$. $\neg P$ is (strictly) preferred to $P \wedge Q$, which in turn is preferred to $P \wedge \neg Q$.

Although the preference relations of case models are qualitative, they correspond to the relations that can be represented by real-valued functions.

Corollary 2 *Let $C \subseteq L$ be finite with elements that are logically consistent, mutually incompatible and different (properties 1, 2 and 3 in the definition of case models). Then the following are equivalent:*

1. *$(C, \geq)$ is a case model;*
2. *$\geq$ is numerically representable, i.e., there is a real valued function v on C such that for all φ and $\psi \in C$, $\varphi \geq \psi$ if and only if $v(\varphi) \geq v(\psi)$.*

The function v can be chosen with only positive values, or even with only positive integer values.

Proof. It is a standard result in order theory that total preorders on finite (or countable) sets are the ones that are representable by a real-valued function [21]. QED

Corollary 3 *Let $C \subseteq L$ be finite with elements that are logically consistent, mutually incompatible and different (properties 1, 2 and 3 in the definition of case models). Then the following are equivalent:*

1. *$(C, \geq)$ is a case model;*
2. *$\geq$ is numerically representable by a probability function p on the algebra generated by C such that for all φ and $\psi \in C$, $\varphi \geq \psi$ if and only if $p(\varphi) \geq p(\psi)$.*

Proof. Pick a representing real-valued function v with only positive values as in the previous corollary, and (for elements of C) define the values of p as those of v divided by the sum of the v-values of all cases; then extend to the algebra generated by C. QED

Next we define arguments. Arguments are from premises $\varphi \in L$ to conclusions $\psi \in L$.

Definition 4 *An* argument *is a pair (φ, ψ) with φ and $\psi \in L$. The sentence φ expresses the argument's premises, the sentence ψ its conclusions, and the sentence $\varphi \wedge \psi$ the* case made *by the argument. Generalizing, a sentence $\chi \in L$ is a* premise *of the argument when $\varphi \models \chi$, a* conclusion *when $\psi \models \chi$, and a* position *in the case made by the argument when $\varphi \wedge \psi \models \chi$. An argument (φ, ψ) is* (properly) presumptive *when $\varphi \not\models \psi$; otherwise* non-presumptive*. An argument (φ, ψ) is an* assumption *when $\models \varphi$, i.e., when its premises are logically tautologous.*

Note our use of the plural for an argument's premises, conclusions and positions. This terminological convention allows us to speak of the premises $\mathtt{p}$ and $\neg\mathtt{q}$ and conclusions $\mathtt{r}$ and $\neg\mathtt{s}$ of the argument $(\mathtt{p} \wedge \neg\mathtt{q}, \mathtt{r} \wedge \neg\mathtt{s})$. Also the convention fits our non-syntactic definitions, where for instance an argument with premise χ also has logically equivalent sentences such as $\neg\neg\chi$ as a premise.

Coherent arguments are defined as arguments that make a case logically implied by a case in the case model.

Definition 5 *Let $(C, \geq)$ be a case model. Then we define, for all φ and $\psi \in L$:*

$(C, \geq) \models (\varphi, \psi)$ *if and only if* $\exists \omega \in C$*:* $\omega \models \varphi \wedge \psi$.

We then say that the argument from φ to ψ is coherent *with respect to the case model. We say that a coherent argument from φ to ψ is* conclusive *when all cases implying the premises also imply the conclusions.*

Example (continued). In the case model of Figure 2, the arguments from $\top$ to $\neg P$ and to P, and from P to Q and to $\neg Q$ are coherent and not conclusive in the sense of this definition. Denoting the case model as $(C, \geq)$, we have $(C, \geq) \models (\top, \neg P)$, $(C, \geq) \models (\top, P)$, $(C, \geq) \models (P, Q)$ and $(C, \geq) \models (P, \neg Q)$. The arguments from a case (in the case model) to itself, such as from $\neg P$ to $\neg P$, or from $P \wedge Q$ to $P \wedge Q$ are conclusive. The argument $(P \vee R, P)$ is also conclusive in this case model, since all $P \vee R$-cases are P-cases. Similarly, $(P \vee R, P \vee S)$ is conclusive.

The notion of presumptive validity considered here is based on the idea that some arguments make a better case than other arguments from the same premises. More precisely, an argument is presumptively valid if there is a case in the case model implying the case made by the argument that is at least as preferred as all cases implying the premises.

Definition 6 *Let $(C, \geq)$ be a case model. Then we define, for all φ and $\psi \in L$:*

$(C, \geq) \models \varphi \rightsquigarrow \psi$ *if and only if* $\exists \omega \in C$*:*

1. $\omega \models \varphi \wedge \psi$*; and*
2. $\forall \omega' \in C$ *: if* $\omega' \models \varphi$*, then* $\omega \geq \omega'$.

We then say that the argument from φ to ψ is (presumptively) valid *with respect to the case model. A presumptively valid argument is* (properly) defeasible*, when it is not conclusive.*

Example (continued). In the case model of Figure 2, the arguments from $\top$ to $\neg P$, and from P to Q are presumptively valid in the sense of this definition. Denoting the case model as $(C, \geq)$, we have formally that $(C, \geq) \models \top \rightsquigarrow \neg P$ and $(C, \geq) \models P \rightsquigarrow Q$. The coherent arguments from $\top$ to P and from P to $\neg Q$ are not presumptively valid in this sense.

Corollary 7 *1. Conclusive arguments are coherent, but there are case models with a coherent, yet inconclusive argument;*
2. *Conclusive arguments are presumptively valid, but there are case models with a presumptively valid, yet inconclusive argument;*
3. *Presumptively valid arguments are coherent, but there are case models with a coherent, yet presumptively invalid argument.*

The next proposition provides key logical properties of this notion of presumptive validity. Many have been studied for nonmonotonic inference relations [13, 14, 27]. Given a case model $(C, \geq)$, we write $\varphi \mathrel{|\!\sim} \psi$ for $(C, \geq) \models \varphi \rightsquigarrow \psi$. We write $C(\varphi)$ for the set $\{\omega \in C \mid \omega \models \varphi\}$.

(LE), for Logical Equivalence, expresses that in a valid argument the premises and the conclusions can be replaced by a classical equivalent (in the sense of $\models$). (Cons), for Consistency, expresses that the conclusions of presumptively valid arguments must be consistent. (Ant), for Antededence, expresses that when certain premises validly imply a conclusion, the case made by the argument is also validly implied by these premises. (RW), for Right Weakening, expresses that when the premises validly imply a composite conclusion also the intermediate conclusions are validly implied. (CCM), for Conjunctive Cautious Monotony, expresses that the case made by a valid argument is still validly implied when an intermediate conclusion is added to the argument's premises. (CCT), for Conjunctive Cumulative Transitivity, is a variation of the related property Cumulative Transitivity property (CT, also known as Cut). (CT)—extensively studied in the literature—has $\varphi \mathrel{|\!\sim} \chi$ instead of $\varphi \mathrel{|\!\sim} \psi \wedge \chi$ as a consequent. The variation is essential in our setting where the (And) property is absent (If $\varphi \mathrel{|\!\sim} \psi$ and $\varphi \mathrel{|\!\sim} \chi$, then $\varphi \mathrel{|\!\sim} \psi \wedge \chi$). Assuming (Ant), (CCT) expresses the validity of chaining valid implication from φ via the case made in the first step $\varphi \wedge \psi$ to the case made in the second step $\varphi \wedge \psi \wedge \chi$. (See [29, 30], introducing (CCT).)

Proposition 8 *Let $(C, \geq)$ be a case model. For all φ, ψ and $\chi \in L$:*

(LE) *If $\varphi \mathrel{|\!\sim} \psi$, $\models \varphi \leftrightarrow \varphi'$ and $\models \psi \leftrightarrow \psi'$, then $\varphi' \mathrel{|\!\sim} \psi'$.*
(Cons) *$\varphi \not\mathrel{|\!\sim} \bot$.*
(Ant) *If $\varphi \mathrel{|\!\sim} \psi$, then $\varphi \mathrel{|\!\sim} \varphi \wedge \psi$.*
(RW) *If $\varphi \mathrel{|\!\sim} \psi \wedge \chi$, then $\varphi \mathrel{|\!\sim} \psi$.*
(CCM) *If $\varphi \mathrel{|\!\sim} \psi \wedge \chi$, then $\varphi \wedge \psi \mathrel{|\!\sim} \chi$.*
(CCT) *If $\varphi \mathrel{|\!\sim} \psi$ and $\varphi \wedge \psi \mathrel{|\!\sim} \chi$, then $\varphi \mathrel{|\!\sim} \psi \wedge \chi$.*

Proof. (LE): Direct from the definition. (Cons): Otherwise there would be an inconsistent element of C, contradicting the definition of a case model. (Ant): When $\varphi \mathrel{|\!\sim} \psi$, there is an ω with $\omega \models \varphi \wedge \psi$ that is $\geq$-maximal in $C(\varphi)$. Then also $\omega \models \varphi \wedge \varphi \wedge \psi$, hence $\varphi \mathrel{|\!\sim} \varphi \wedge \psi$. (RW): When $\varphi \mathrel{|\!\sim} \psi \wedge \chi$, there is an $\omega \in C$ with $\omega \models \varphi \wedge \psi \wedge \chi$ that is maximal in $C(\varphi)$. Since then also $\omega \models \varphi \wedge \psi$, we find $\varphi \mathrel{|\!\sim} \psi$. (CCM): By the assumption, we have an $\omega \in C$ with $\omega \models \varphi \wedge \psi \wedge \chi$ that is maximal in $C(\varphi)$. Since $C(\varphi \wedge \psi) \subseteq C(\varphi)$, ω is also maximal in $C(\varphi \wedge \psi)$, and we find $\varphi \wedge \psi \mathrel{|\!\sim} \chi$. (CCT): Assuming $\varphi \mathrel{|\!\sim} \psi$, there is an $\omega \in C$ with $\omega \models \varphi \wedge \psi$, maximal in $C(\varphi)$. Assuming also $\varphi \wedge \psi \mathrel{|\!\sim} \chi$, there is an $\omega' \in C$ with $\omega \models \varphi \wedge \psi \wedge \chi$, maximal in $C(\varphi \wedge \psi)$. Since $\omega \in C(\varphi \wedge \psi)$, we find $\omega' \geq \omega$. By transitivity of $\geq$, and the maximality of ω in $C(\varphi)$, we therefore have that ω' is maximal in $C(\varphi)$. As a result, $\varphi \mathrel{|\!\sim} \psi \wedge \chi$. QED

We speak of coherent premises when the argument from the premises to themselves is coherent. The following proposition provides some equivalent characterizations of coherent premises.

Proposition 9 *Let $(C, \geq)$ be a case model. The following are equivalent, for all $\varphi \in L$:*

1. $\varphi \mathrel{|\!\sim} \varphi$*;*
2. $\exists \omega \in C : \omega \models \varphi$ *and* $\forall \omega' \in C$*: If* $\omega' \models \varphi$*, then* $\omega \geq \omega'$*;*
3. $\exists \omega \in C : \varphi \mathrel{|\!\sim} \omega$.
4. $\exists \omega \in C : \omega \models \varphi$.

Proof. 1 and 2 are equivalent by the definition of $\mathrel{|\!\sim}$. Assume 2. Then there is a $\geq$-maximal element ω of $C(\varphi)$. By the definition of $\mathrel{|\!\sim}$,

then $\varphi \mid\!\sim \omega$; proving 3. Assume 3. Then there is a $\geq$-maximal element ω' of $C(\varphi)$ with $\omega' \models \varphi \wedge \omega$. For this ω' also $\omega' \models \varphi$, showing 2. 4 logically follows from 2. 4 implies 2 since L is a language that generated by finitely many propositional constants. QED

Corollary 10 *Let $(C, \geq)$ be a case model. Then all coherent arguments have coherent premises and all presumptively valid arguments have coherent premises.*

We saw that, in the present approach, premises are coherent when they are logically implied by a case in the case model. As a result, generalisations of coherent premises are again coherent; cf. the following corollary.

Corollary 11 *Let $(C, \geq)$ be a case model. Then:*

If $\varphi \mid\!\sim \varphi$ and $\varphi \models \psi$, then $\psi \mid\!\sim \psi$.

We now consider some properties that use a subset L^* of the language L. The set L^* consists of the logical combinations of the cases of the case model using negation, conjunction and logical equivalence (cf. the algebra underlying probability functions [21]). L^* is the set of *case expressions* associated with a case model.

(Coh), for Coherence, expresses that coherent premises correspond to a consistent case expression implying the premises. (Ch), for Choice, expresses that, given two coherent case expressions, at least one of three options follows validly: the conjunction of the case expression, or the conjunction of one of them with the negation of the other. (OC), for Ordered Choice, expresses that preferred choices between case expressions are transitive. Here we say that a case expression is a *preferred choice* over another, when the former follows validly from the disjunction of both.

Definition 12 *Let $(C, \geq)$ be a case model, $\varphi \in L$, and $\omega \in C$. Then ω expresses a* preferred case *of φ if and only if $\varphi \mid\!\sim \omega$.*

Proposition 13 *Let $(C, \geq)$ be a case model, and $L^* \subseteq L$ the closure of C under negation, conjunction and logical equivalence. Writing $\mid\!\sim^*$ for the restriction of $\mid\!\sim$ to L^*, we have, for all φ, ψ and $\chi \in L^*$:*

(Coh) $\varphi \mid\!\sim \varphi$ *if and only if* $\exists \varphi^* \in L^*$ *with* $\varphi^* \not\models \bot$ *and* $\varphi^* \models \varphi$*;*
(Ch) *If* $\varphi \mid\!\sim^* \varphi$ *and* $\psi \mid\!\sim^* \psi$*, then* $\varphi \vee \psi \mid\!\sim^* \neg\varphi \wedge \psi$ *or* $\varphi \vee \psi \mid\!\sim^* \varphi \wedge \psi$ *or* $\varphi \vee \psi \mid\!\sim^* \varphi \wedge \neg\psi$*;*
(OC) *If* $\varphi \vee \psi \mid\!\sim^* \varphi$ *and* $\psi \vee \chi \mid\!\sim^* \psi$*, then* $\varphi \vee \chi \mid\!\sim^* \varphi$*.*

Proof. (Coh): By Proposition 9, $\varphi \mid\!\sim \varphi$ if and only if there is an $\omega \in C$ with $\omega \models \varphi$. The property (Coh) follows since $C \subseteq L^*$ and, for all consistent $\varphi^* \in L^*$, there is an $\omega \in C$ with $\omega \models \varphi^*$.
(Ch): Consider sentences φ and $\psi \in L^*$ with $\varphi \mid\!\sim^* \varphi$ and $\psi \mid\!\sim^* \psi$. Then, by Corollary 11, $\varphi \vee \psi \mid\!\sim \varphi \vee \psi$. By Proposition 9, there is an $\omega \in C$, with $\omega \models \varphi \vee \psi$. The sentences φ and ψ are elements of L^*, hence also the sentences $\varphi \wedge \neg\psi$, $\varphi \wedge \psi$ and $\neg\varphi \wedge \psi \in L^*$. All are logically equivalent to disjunctions of elements of C (possibly the empty disjunction, logically equivalent to $\bot$). Since $\omega \models \varphi \vee \psi$, $\models \varphi \vee \psi \leftrightarrow (\varphi \wedge \neg\psi) \vee (\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi)$, and the elements of C are mutually incompatible, we have $\omega \models \varphi \wedge \neg\psi$ or $\omega \models \varphi \wedge \psi$ or $\omega \models \neg\varphi \wedge \psi$. By Proposition 9, it follows that $\varphi \vee \psi \mid\!\sim^* \neg\varphi \wedge \psi$ or $\varphi \vee \psi \mid\!\sim^* \varphi \wedge \psi$ or $\varphi \vee \psi \mid\!\sim^* \varphi \wedge \neg\psi$.
(OC): By $\varphi \vee \psi \mid\!\sim^* \varphi$, there is an $\omega \models \varphi$ maximal in $C(\varphi \vee \psi)$. By $\psi \vee \chi \mid\!\sim^* \psi$, there is an $\omega' \models \psi$ maximal in $C(\psi \vee \chi)$. Since $\omega \models \varphi$, $\omega \in C(\varphi \vee \chi)$. Since $\omega' \models \psi$, $\omega' \in C(\varphi \vee \psi)$, hence $\omega \geq \omega'$. Hence ω is maximal in $C(\varphi \vee \chi)$, hence $\varphi \vee \chi \mid\!\sim \varphi$. Since $\chi \in L^*$, $\varphi \vee \chi \mid\!\sim^* \varphi$. QED

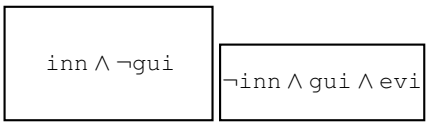


Figure 3. A case model for presumption

4 A formal analysis of some key concepts

We now use the formalism of case models and presumptive validity above for a discussion of some key concepts associated with the argumentative, scenario and probabilistic analysis of evidential reasoning.

4.1 Arguments

In an argumentative analysis, it is natural to classify arguments with respect to the nature of the support their premises give their conclusions. We already defined non-presumptive and (properly) presumptive arguments (Definition 4), and—with respect to a case model—presumptively valid and (properly) defeasible arguments (Definition 6). We illustrate these notions in an example about the presumption of innocence.

Let inn denote that a suspect is innocent, and gui that he is guilty. Then the argument $(\texttt{inn}, \neg\texttt{gui})$ is (properly) presumptive, since $\texttt{inn} \not\models \neg\texttt{gui}$. The argument $(\texttt{inn} \wedge \neg\texttt{gui}, \neg\texttt{gui})$ is non-presumptive, since $\texttt{inn} \wedge \neg\texttt{gui} \models \neg\texttt{gui}$.

Presumptive validity and (proper) defeasibility are illustrated using a case model. Consider the case model with two cases $\texttt{inn} \wedge \neg\texttt{gui}$ and $\neg\texttt{inn} \wedge \texttt{gui} \wedge \texttt{evi}$ with the first case preferred to the second (Figure 3; the surface of the cases measures their preference). Here evi denotes evidence for the suspect's guilt. Then the (properly) presumptive argument $(\texttt{inn}, \neg\texttt{gui})$ is presumptively valid with respect to this case model since the conclusion $\neg\texttt{gui}$ follows in the case $\texttt{inn} \wedge \neg\texttt{gui}$ that is a preferred case of the premise inn. The argument is conclusive since there are no other cases implying inn. The argument $(\top, \texttt{inn})$—in fact an assumption now that its premises are tautologous—is presumptively valid since inn follows in the preferred case $\texttt{inn} \wedge \neg\texttt{gui}$. This shows that the example represents what is called the presumption of innocence, when there is no evidence. This argument is (properly) defeasible since in the other case of the argument's premises the conclusion does not follow. In fact, the argument $(\texttt{evi}, \texttt{inn})$ is not coherent since there is no case in which both evi and inn follow. The argument $(\texttt{evi}, \texttt{gui})$ is presumptively valid, even conclusive.

In argumentative analyses, different kinds of argument attack are considered. John Pollock made the famous distinction between two kinds of—what he called—argument defeaters [17, 18]. A rebutting defeater is a reason for a conclusion that is the opposite of the conclusion of the attacked argument, whereas an undercutting defeater is a reason that attacks not the conclusion itself, but the connection between reason and conclusion. Joseph Raz made a related famous distinction of exclusionary reasons that always prevail, independent of the strength of competing reasons [19] (see also [20]).

We propose the following terminology.

Definition 14 *Let $(C, \geq)$ be a case model, and (φ, ψ) a presumptively valid argument. Then circumstances χ are* undercutting *when $(\varphi \wedge \chi, \psi)$ is not presumptively valid. Undercutting circumstances are* rebutting *when $(\varphi \wedge \chi, \neg\psi)$ is presumptively valid; otherwise they are* properly undercutting*. Undercutting circumstances are* excluding *when $(\varphi \wedge \chi, \psi)$ is not coherent.*

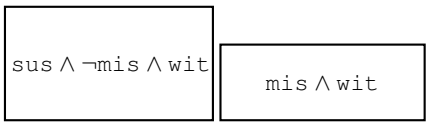


Figure 4. A case model for proper undercutting

Continuing the example of the case model illustrated in Figure 3, we find the following. The circumstances evi undercut the presumptively valid argument $(\top, \texttt{inn})$ since $(\texttt{evi}, \texttt{inn})$ is not presumptively valid. In fact, these circumstances are excluding since $(\texttt{evi}, \texttt{inn})$ is not coherent. The circumstances are also rebutting since the argument for the opposite conclusion $(\texttt{evi}, \neg\texttt{inn})$ is presumptively valid.

Proper undercutting can be illustrated with an example about a lying witness. Consider a case model with these two cases:

1: $\texttt{sus} \wedge \neg\texttt{mis} \wedge \texttt{wit}$
2: $\texttt{mis} \wedge \texttt{wit}$

In the cases, there is a witness testimony (wit) that the suspect was at the crime scene (sus). In Case 1, the witness was not misguided ($\neg$mis), in Case 2 he was. In Case 1, the suspect was indeed at the crime scene; in Case 2, the witness was misguided and it is unspecified whether the suspect was at the crime scene or not. In the case model, Case 1 is preferred to Case 2 (Figure 4), representing that witnesses are usually not misguided.

Since Case 1 is a preferred case of wit, the argument $(\texttt{wit}, \texttt{sus})$ is presumptively valid: the witness testimony provides a presumptively valid argument for the suspect having been at the crime scene. The argument's conclusion can be strengthened to include that the witness was not misguided. Formally, this is expressed by saying that $(\texttt{wit}, \texttt{sus} \wedge \neg\texttt{mis})$ is a presumptively valid argument. There are circumstances undercutting the argument $(\texttt{wit}, \texttt{sus})$, namely when the witness was misguided after all (mis). This can be seen by considering that Case 2 is the only case in which $\texttt{wit} \wedge \texttt{mis}$ follows, hence is preferred. Since sus does not follow in Case 2, the argument $(\texttt{wit} \wedge \texttt{mis}, \texttt{sus})$ is not presumptively valid. The misguidedness is not rebutting, hence properly undercutting since $(\texttt{wit} \wedge \texttt{mis}, \neg\texttt{sus})$ is not presumptively valid. The misguidedness is excluding since the argument $(\texttt{wit} \wedge \texttt{mis}, \texttt{sus})$ is not even coherent.

Arguments can typically be chained, namely when the conclusion of one is a premise of another. For instance when there is evidence (evi) that a suspect is guilty of a crime (gui), the suspect's guilt can be the basis of punishing the suspect (pun). For both steps there are typical undercutting circumstances. The step from the evidence to guilt is blocked when there is an alibi (ali), and the step from guilt to punishing is blocked when there are grounds of justification (jus), such as force majeure. A case model with three cases can illustrate such chaining:

1: $\texttt{pun} \wedge \texttt{gui} \wedge \texttt{evi}$
2: $\neg\texttt{pun} \wedge \texttt{gui} \wedge \texttt{evi} \wedge \texttt{jus}$
3: $\neg\texttt{gui} \wedge \texttt{evi} \wedge \texttt{ali}$

In the case model, Case 1 is preferred to Case 2 and Case 3, modeling that the evidence typically leads to guilt and punishing, unless there are grounds for justification (Case 2) or there is an alibi (Case 3). Cases 2 and 3 are preferentially equivalent.

In this case model, the following arguments are presumptively valid:

1: $(\texttt{evi}, \texttt{gui})$
2: $(\texttt{gui}, \texttt{pun})$
3: $(\texttt{evi}, \texttt{gui} \wedge \texttt{pun})$

Arguments 1 and 3 are presumptively valid since Case 1 is the preferred case among those in which evi follows; Argument 2 is since Case 1 is the preferred case among those in which gui follows. By chaining arguments 1 and 2, the case for $\texttt{gui} \wedge \texttt{pun}$ can be based on the evidence evi as in Argument 3.

The following arguments are not presumptively valid in this case model:

4: $(\texttt{evi} \wedge \texttt{ali}, \texttt{gui})$
5: $(\texttt{gui} \wedge \texttt{jus}, \texttt{pun})$

This shows that Arguments 1 and 2 are undercut by circumstances ali and jus, respectively. As expected, chaining these arguments fails under both of these circumstances, as shown by the fact that these two arguments are not presumptively valid:

6: $(\texttt{evi} \wedge \texttt{ali}, \texttt{gui} \wedge \texttt{pun})$
7: $(\texttt{evi} \wedge \texttt{jus}, \texttt{gui} \wedge \texttt{pun})$

But the step to guilt can be made when there are grounds for justification. Formally, this can be seen by the presumptive validity of this argument:

8: $(\texttt{evi} \wedge \texttt{jus}, \texttt{gui})$

4.2 Scenarios

In the literature on scenario analyses, several notions are used in order to analyze the 'quality' of the scenarios considered. Three notions are prominent: a scenario's consistency, a scenario's completeness and a scenario's plausibility [16, 37]. In this literature, these notions are part of an informally discussed theoretical background, having prompted recent work in AI & Law on formalizing these notions [3, 33, 36]. A scenario is consistent when it does not contain contradictions. For instance, a suspect cannot be both at home and at the crime scene. A scenario is complete when all relevant elements are in the scenario. For instance, a murder scenario requires a victim, an intention and premeditation. A scenario is plausible when it fits commonsense knowledge about the world. For instance, in a murder scenario, a victim's death caused by a shooting seems a plausible possibility. We now propose a formal treatment of these notions using the formalism presented.

The consistency of a scenario can simply be taken to correspond to logical consistency. A more interesting, stronger notion of scenario-consistency uses the world knowledge takes represented in a case model and defines a scenario as scenario-consistent when it is a logically consistent coherent assumption. Formally, writing S for the scenario, S is scenario-consistent when S is logically consistent and the argument $(\top, S)$ is coherent, i.e., there is a case in the case model logically implying S.

The completeness of a scenario can here be defined using a notion of maximally specific conclusions, as follows.

Definition 15 *Let $(C, \geq)$ be a case model, and (φ, ψ) a presumptively valid argument. Then the case made by the argument (i.e., $\varphi \wedge \psi$) is an* extension *of φ when there is no presumptively valid argument from φ that makes a case that is logically more specific.*

For instance, consider a case model in which the case $\texttt{vic} \wedge \texttt{int} \wedge \texttt{pre} \wedge \texttt{evi}$ is a preferred case of $\texttt{evi}$. The case expresses a situation in which there is evidence ($\texttt{evi}$) for a typical murder: there is a victim ($\texttt{vic}$), there was the intention to kill ($\texttt{int}$), and there was premeditation ($\texttt{pre}$). In such a case model, this case is an extension of the evidence $\texttt{evi}$. A scenario can now be considered complete with respect to certain evidence when the scenario conjoined with the evidence is its own extension. In the example, the sentence $\texttt{vic} \wedge \texttt{int} \wedge \texttt{pre}$ is a complete scenario given $\texttt{evi}$ as the scenario conjoined with the evidence is its own extension. The sentence $\texttt{vic} \wedge \texttt{int}$ is not a complete scenario given $\texttt{evi}$, as the extension of $\texttt{vic} \wedge \texttt{int} \wedge \texttt{evi}$ also implies $\texttt{pre}$.

A scenario can be treated as plausible (given a case model) when it is a presumptively valid conclusion of the evidence. Continuing the example, the complete scenario $\texttt{vic} \wedge \texttt{int} \wedge \texttt{pre}$ is then plausible given $\texttt{evi}$, but also subscenarios such as $\texttt{vic} \wedge \texttt{int}$ (leaving the premeditation unspecified) and $\texttt{int} \wedge \texttt{pre}$ (with no victim, only intention and premeditation). This notion of a scenario's plausibility depends on the evidence, in contrast with the mentioned literature [16, 37], where plausibility is treated as being independent from the evidence. The present proposal includes an evidence-independent notion of plausibility, by considering a scenario as plausible—independent of the evidence—when it is plausible given no evidence, i.e., when the scenario is a presumptively valid assumption. In the present setting, plausibility can be connected to the preference ordering on cases given the evidence, when scenarios are complete. A complete scenario is than more plausible than another given the evidence when the former is preferred to the latter.

4.3 Probabilities

The literature on the probabilistic analysis of reasoning with evidence uses the probability calculus as formal background. A key formula is the well-known Bayes' theorem, stating that for events H and E the following relation between probabilities holds:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \cdot \Pr(H)$$

Thinking of H as a hypothesis and E as evidence, here the posterior probability $\Pr(H|E)$ of the hypothesis given the evidence can be computed by multiplying the prior probability $\Pr(H)$ and the Bayes factor $\Pr(E|H)/\Pr(E)$.

We saw that the preferences of our case models are exactly those that can be realized by probability functions over the cases in the model (Corollary 3). Given a realization of a case model, key concepts defined in terms of the case model translate straightforwardly to the probabilistic setting. For instance, a preferred case (given certain premises) has maximal probability (conditional on these premises) among the cases from which the premises follow. Also the premises provide a conclusive argument for a case if there is exactly one case from which the premises follow, hence if the probability of the case given the premises is equal to 1. Also, clearly, Bayes' theorem holds for any such probabilistic realization of our case models.

A formula that is especially often encountered in the literature on evidential reasoning is the following odds version of Bayes' theorem:

$$\frac{\Pr(H|E)}{\Pr(\neg H|E)} = \frac{\Pr(E|H)}{\Pr(E|\neg H)} \cdot \frac{\Pr(H)}{\Pr(\neg H)}$$

Here the posterior odds $\Pr(H|E)/\Pr(\neg H|E)$ of the hypothesis given the evidence is found by multiplying the prior odds $\Pr(H)/\Pr(\neg H)$ with the likelihood ratio $\Pr(E|H)/\Pr(E|\neg H)$. This formula is important since the likelihood ratio can sometimes be estimated, for instance in the case of DNA evidence. In fact, it is a key lesson in probabilistic approaches to evidential reasoning that the evidential value of evidence, as measured by a likelihood ratio, does not by itself determine the posterior probability of the hypothesis considered. As the formula shows, the prior probability of the hypothesis is needed to determine the posterior probability given the likelihood ratio. Just as Bayes' theorem, the likelihood ratio obtains in a probabilistic realization of a case model in our sense.

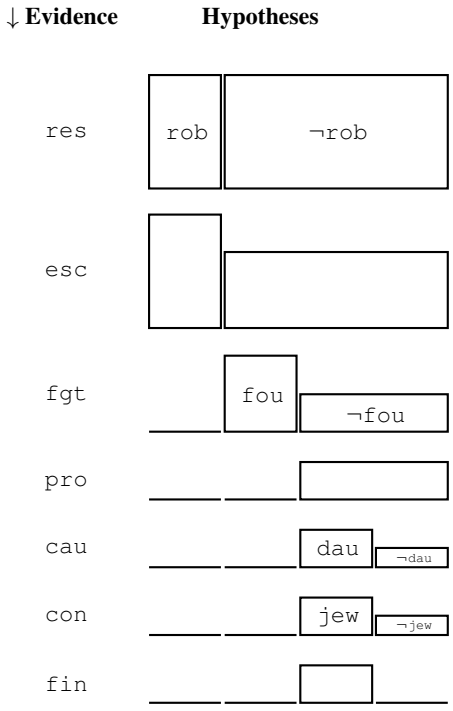


Figure 5. Example: Hitchcock's 'To Catch A Thief'

5 Example: Alfred Hitchcock's 'To Catch A Thief'

As an example of the development of evidential reasoning in which gradually information is collected, we discuss the crime investigation story that is the backbone of Alfred Hitchcock's 'To Catch A Thief', otherwise—what Hitchcock himself referred to as—a lightweight story about a French Riviera love affair, starring Grace Kelly and Cary Grant. In the film, Grant plays a former robber Robie, called 'The Cat' because of his spectacular robberies, involving the climbing of high buildings. At the beginning of the film, new 'The Cat'-like thefts have occurred. Because of this resemblance with Robie's style (the first evidence considered, denoted in what follows as $\texttt{res}$), the police consider the hypothesis that Robie is again the thief ($\texttt{rob}$), and also that he is not ($\neg\texttt{rob}$). Figure 5 provides a graphical representation of the investigation. The first row shows the situation after the first evidence $\texttt{res}$, mentioned on the left side of the figure, with the two hypothical conclusions $\texttt{rob}$ and $\neg\texttt{rob}$ represented as rectangles. A rectangle's height suggests the strength of the argument from the accumulated evidence to the hypothesis. Here the arguments from $\texttt{res}$ to $\texttt{rob}$ and $\neg\texttt{rob}$ are of comparable strength.

When the police confront Robie with the new thefts, he escapes with the goal to catch the real thief. By this second evidence ($\texttt{esc}$), the hypothesis $\texttt{rob}$ becomes more strongly supported than its oppo-

site ¬rob. In the figure, the second row indicates the situation after the two pieces of evidence are available. As indicated by the rectangles of different heights, the argument from the accumulated evidence res∧esc to rob is stronger than that from the same premises to ¬rob. Rectangles in a column in the figure represent corresponding hypotheses. Sentences shown in a corresponding hypothesis in a higher row are not repeated.

Robie sets a trap for the real thief, resulting in a nightly fight on the roof with Foussard who falls and dies (fgt). The police consider this strong evidence for the hypothesis that Foussard is the thief (fou), but not conclusive so also the opposite hypothesis is considered coherent (¬fou). In the figure (third row marked fgt) the hypothesis ¬rob is split into two hypotheses: one rectangle representing ¬rob∧fou, the other ¬rob∧¬fou. With the accumulated evidence res∧esc ∧fgt as premises, the hypothesis ¬rob∧fou is more strongly supported than the hypothesis ¬rob∧¬fou. The police no longer believe that Robie is the thief. This is indicated by the line on the left of the third row in the figure. The premises res∧esc ∧fgt do not provide support for the hypothesis rob; or, in the terminology of this paper: the argument from premises res∧esc ∧fgt to conclusion rob is not coherent.

Robie points out that Foussard cannot be the new incarnation of 'The Cat', as he had a prosthetic wooden leg (pro). In other words, the argument from res∧esc ∧fgt∧pro to ¬rob∧fou is not coherent. (Cf. the second line in the fourth row of the figure, corresponding to the hypothesis that Foussard is the thief.)

Later in the film, Foussard's daughter is caught in the act (cau), providing very strong support for the hypothesis that the daughter is the new cat (dau). The argument from res∧esc ∧fgt∧pro∧cau to dau is stronger than to ¬dau.

In her confession (con), Foussard's daughter explains where the jewelry stolen earlier can be found, adding some specific information to the circumstances of her crimes (jew). The argument from res∧esc ∧fgt∧pro∧cau ∧con to dau ∧ jew is stronger than to ¬dau ∧ ¬jew.

The police find the jewelry at the indicated place (fin) and there is no remaining doubt about the hypothesis that Foussard's daughter is the thief. The argument from res∧esc ∧fgt∧pro∧cau∧con ∧fin to ¬dau∧ ¬jew is incoherent, as indicated by the line on the right of the bottom row of the figure. In the only remaining hypothesis, Foussard's daughter is the thief, and not Robie, not Foussard. In other words, the argument from res∧esc ∧fgt∧pro∧cau∧con ∧jew to ¬rob∧¬fou∧dau is conclusive.

We can use the constructions of the representation theorem to develop a case model representing the arguments discussed in the example. We distinguish 7 cases, as follows:

1. rob
   ∧ res∧esc
2. ¬rob∧fou
   ∧ res∧esc∧fgt
3. ¬rob∧¬fou∧dau∧jew
   ∧ res∧esc∧fgt∧pro∧cau∧con∧fin
4. ¬rob∧¬fou∧¬dau∧¬jew
   ∧ res∧esc∧fgt∧pro∧cau∧con
5. ¬rob
   ∧res∧¬esc
6. ¬rob∧¬fou
   ∧res∧esc∧¬fgt
7. ¬rob∧¬fou∧¬dau
   ∧res∧esc∧fgt∧pro∧¬cau

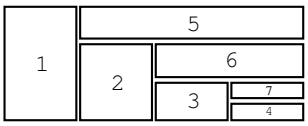


Figure 6. Case model for the example

Cases 1 to 4 are found as follows. First the properties of the four main hypotheses are accumulated (rob, ¬rob∧fou, ¬rob∧¬fou∧dau∧jew, ¬rob∧¬fou∧¬dau∧¬jew). Then these are conjoined with the maximally specific accumulated evidence that provide a coherent argument for them (res∧esc, res∧esc∧fgt, res∧esc∧fgt∧pro∧cau∧con∧fin, res∧esc∧fgt∧pro∧cau∧con). The cases 5 to 7 complete the case model. Case 5 is the hypothetical case that Robie is not the thief, that there is resemblance, and the Robie does not escape. In case 6, Robie and Foussard are not the thieves, and there is no fight. In case 7, Robie, Foussard and his daughter are not the thieves, and she is not caught in the act. Note that the cases are consistent and mutually exclusive.

Figure 6 shows the 7 cases of the model. The sizes of the rectangles represent the preferences. The preference relation has the following equivalence classes, ordered from least preferred to most preferred:

1. Cases 4 and 7;
2. Case 3;
3. Cases 2 and 6;
4. Cases 1 and 5.

The discussion of the arguments, their coherence, conclusiveness and validity presented semi-formally above fits this case model. For instance, the argument from the evidential premises res∧esc to the hypothesis rob is presumptively valid in this case model since Case 1 is the only case implying the case made by the argument. It is not conclusive since also the argument from these same premises to ¬rob is coherent. The latter argument is not presumptively valid since all cases implying the premises have lower preference than Case 1. The argument from res∧esc∧fgt to rob is incoherent as there is no case in which the premises and the conclusion follow. Also arguments that do not start from evidential premises can be evaluated. For instance, the argument from the premise (not itself evidence) dau to jew is conclusive since in the only case implying the premises (Case 3) the conclusion follows. Finally we find the conclusive argument from premises res∧esc∧fgt∧pro∧cau∧con∧jew to conclusion ¬rob∧¬fou∧dau∧jew (only Case 3 implies the premises), hence also to dau.

6 Concluding remarks

In this paper, we have discussed correct reasoning with evidence using three analytic tools: arguments, scenarios and probabilities. We proposed a formalism in which the presumptive validity of arguments is defined in terms of case models, and studied some properties (Section 3). We discussed key concepts in the argumentative, scenario and probabilistic analysis of reasoning with evidence in terms of the formalism (Section 4). An example of the gradual development of evidential reasoning was provided in Section 5.

This work builds on a growing literature aiming to formally connect the three analytic tools of arguments, scenarios and probabilities. In a discussion of the anchored narratives theory by Crombag, Wagenaar and Van Koppen [37], it was shown how argumentative

notions were relevant in their scenario analyses [28]. Bex [3, 5] has provided a hybrid model connecting arguments and scenarios, and has worked on the further integration of the two tools [4, 6]. Connections between arguments and probabilities have been studied by Hepler, Dawid and Leucari [10] combining object-oriented modeling and Bayesian networks. Fenton, Neil and Lagnado continued this work by developing representational idioms for the modeling of evidential reasoning in Bayesian networks [9]. Inspired by this research, Vlek developed scenario idioms for the design of evidential Bayesian networks containing scenarios [35], and Timmer showed how argumentative information can be extracted from a Bayesian network [25]. Keppens and Schafer [12] studied the knowledge-based generation of hypothetical scenarios for reasoning with evidence, later developed further in a decision support system [22].

The present paper continues from an integrated perspective on arguments, scenarios and probabilities [32]. In the present paper, that integrated perspective is formally developed (building on ideas in [31]) using case models and discussing key concepts used in argumentative, scenario and probabilistic analyses. Interestingly, our case models and their preferences are qualitative in nature, while the preferences correspond exactly to those that can be numerically and probabilistically realized. As such, the present formal tools combine a non-numeric and numeric perspective (cf. [32]'s 'To Catch A Thief With and Without Numbers'). Also the present work does not require modeling evidential reasoning in terms of full probability functions, as is the case in Bayesian network approaches. In this way, the well-known problem of needing to specify more numbers than are reasonably available is addressed. Also whereas the causal interpretation of Bayesian networks is risky [7], our case models come with formal definitions of arguments and their presumptive validity.

By the present and related studies, we see a gradual clarification of how arguments, scenarios and probabiities all have their specific useful place in the analysis of evidential reasoning. In this way, it seems ever less natural to choose between the three kinds of tools, and ever more so to use each of them when practically applicable.

ACKNOWLEDGEMENTS

The research reported in this paper has been performed in the context of the project 'Designing and Understanding Forensic Bayesian Networks with Arguments and Scenarios', funded in the NWO Forensic Science program (`http://www.ai.rug.nl/~verheij/nwofs/`).

REFERENCES

[1] T. Anderson, D. Schum, and W. Twining, *Analysis of Evidence. 2nd Edition*, Cambridge University Press, Cambridge, 2005.

[2] W. L. Bennett and M. S. Feldman, *Reconstructing Reality in the Courtroom*, London: Tavistock Feldman, 1981.

[3] F. J. Bex, *Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory*, Springer, Berlin, 2011.

[4] F. J. Bex, 'An integrated theory of causal scenarios and evidential arguments', in *Proceedings of the 15th International Conference on Artificial Intelligence and Law (ICAIL 2015)*, 13–22, ACM Press, New York, (2015).

[5] F. J. Bex, P. J. van Koppen, H. Prakken, and B. Verheij, 'A hybrid formal theory of arguments, stories and criminal evidence', *Artificial Intelligence and Law*, **18**, 1–30, (2010).

[6] F. J. Bex and B. Verheij, 'Legal stories and the process of proof', *Artificial Intelligence and Law*, **21**(3), 253–278, (2013).

[7] A. P. Dawid, 'Beware of the DAG!', in *JMLR Workshop and Conference Proceedings: Volume 6. Causality: Objectives and Assessment (NIPS 2008 Workshop)*, eds., I. Guyon, D. Janzing, and B. Schölkopf, 59–86, jmlr.org, (2010).

[8] *Evidence, Inference and Enquiry*, eds., A. P. Dawid, W. Twining, and M. Vasiliki, Oxford University Press, Oxford, 2011.

[9] N. E. Fenton, M. D. Neil, and D. A. Lagnado, 'A general structure for legal arguments about evidence using Bayesian Networks', *Cognitive Science*, **37**, 61–102, (2013).

[10] A. B. Hepler, A. P. Dawid, and V. Leucari, 'Object-oriented graphical representations of complex patterns of evidence', *Law, Probability and Risk*, **6**(1–4), 275–293, (2007).

[11] *Legal Evidence and Proof: Statistics, Stories, Logic (Applied Legal Philosophy Series)*, eds., H. Kaptein, H. Prakken, and B. Verheij, Ashgate, Farnham, 2009.

[12] J. Keppens and B. Schafer, 'Knowledge based crime scenario modelling', *Expert Systems with Applications*, **30**(2), 203–222, (2006).

[13] S. Kraus, D. Lehmann, and M. Magidor, 'Nonmonotonic reasoning, preferential models and cumulative logics', *Artificial Intelligence*, **44**, 167–207, (1990).

[14] D. Makinson, 'General patterns in nonmonotonic reasoning', in *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Nonmonotonic Reasoning and Uncertain Reasoning*, eds., D. M. Gabbay, C. J. Hogger, and J. A. Robinson, 35–110, Clarendon Press, Oxford, (1994).

[15] J. Pearl, *Causality: Models, Reasoning and Inference. Second Edition*, Cambridge University Press, Cambridge, 2000/2009.

[16] N. Pennington and R. Hastie, 'Reasoning in explanation-based decision making', *Cognition*, **49**(1–2), 123–163, (1993).

[17] J. L. Pollock, 'Defeasible reasoning', *Cognitive Science*, **11**(4), 481–518, (1987).

[18] J. L. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person*, The MIT Press, Cambridge (Massachusetts), 1995.

[19] J. Raz, *Practical Reason and Norms*, Princeton University Press, Princeton (New Jersey), 1990.

[20] H. S. Richardson, 'Moral reasoning', in *The Stanford Encyclopedia of Philosophy*, ed., E. N. Zalta, Stanford University, (2013).

[21] F. S. Roberts, *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Cambridge University Press, Cambridge, 1985.

[22] Q. Shen, J. Keppens, C. Aitken, B. Schafer, and M. Lee, 'A scenario-driven decision support system for serious crime investigation', *Law, Probability and Risk*, **5**, 87–117, (2006).

[23] F. Taroni, C. Champod, and P. Margot, 'Forerunners of Bayesianism in early forensic science', *Jurimetrics*, **38**, 183–200, (1998).

[24] S. T. Timmer, J. J. Meyer, H. Prakken, S. Renooij, and B. Verheij, 'Capturing critical questions in Bayesian network fragments. legal knowledge and information systems', in *Legal Knowledge and Information Systems: JURIX 2015: The Twenty-Eighth Annual Conference*, ed., A. Rotolo, 173–176, IOS Press, Amsterdam, (2015).

[25] S. T. Timmer, J. J. Meyer, H. Prakken, S. Renooij, and B. Verheij, 'Explaining Bayesian Networks using argumentation', in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conference, ECSQARU 2015, Compigne, France, July 15-17, 2015. Proceedings*, 83–92, Springer, Berlin, (2015).

[26] S. E. Toulmin, *The Uses of Argument*, Cambridge University Press, Cambridge, 1958.

[27] J. van Benthem, 'Foundations of conditional logic', *Journal of Philosophical Logic*, **13**, 303–349, (1984).

[28] B. Verheij, 'Dialectical argumentation as a heuristic for courtroom decision making', in *Rationality, Information and Progress in Law and Psychology. Liber Amicorum Hans F. Crombag*, eds., P. J. van Koppen and N. Roos, 203–226, Metajuridica Publications, (2000).

[29] B. Verheij, 'Argumentation and rules with exceptions', in *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, eds., B. Baroni, F. Cerutti, M. Giacomin, and G. R. Simari, 455–462, IOS Press, Amsterdam, (2010).

[30] B. Verheij, 'Jumping to conclusions. a logico-probabilistic foundation for defeasible rule-based arguments', in *13th European Conference on Logics in Artificial Intelligence, JELIA 2012. Toulouse, France, September 2012. Proceedings (LNAI 7519)*, eds., L. Fariñas del Cerro, A. Herzig, and J. Mengin, 411–423, Springer, Berlin, (2012).

[31] B. Verheij, 'Arguments and their strength: Revisiting Pollock's anti-probabilistic starting points', in *Computational Models of Argument. Proceedings of COMMA 2014*, eds., S. Parsons, N. Oren, C. Reed, and F. Cerutti, 433–444, IOS Press, Amsterdam, (2014).

[32] B. Verheij, 'To catch a thief with and without numbers: Arguments,

scenarios and probabilities in evidential reasoning', *Law, Probability and Risk*, **13**, 307–325, (2014).
[33] B. Verheij and F. J. Bex, 'Accepting the truth of a story about the facts of a criminal case', in *Legal Evidence and Proof: Statistics, Stories, Logic*, eds., H. Kaptein, H. Prakken, and B. Verheij, 161–193, Ashgate, Farnham, (2009).
[34] B. Verheij, F. J. Bex, S. T. Timmer, C. S. Vlek, J. J. Meyer, S. Renooij, and H. Prakken, 'Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning', *Law, Probability and Risk*, **15**, 35–70, (2016).
[35] C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij, 'Building Bayesian Networks for legal evidence with narratives: a case study evaluation', *Artifical Intelligence and Law*, **22**(4), 375–421, (2014).
[36] C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij, 'Representing the quality of crime scenarios in a Bayesian network', in *Legal Knowledge and Information Systems: JURIX 2015: The Twenty-Eighth Annual Conference*, ed., A. Rotolo, 131–140, IOS Press, Amsterdam, (2015).
[37] W. A. Wagenaar, P. J. van Koppen, and H. F. M. Crombag, *Anchored Narratives. The Psychology of Criminal Evidence*, Harvester Wheatsheaf, London, 1993.
[38] J. H. Wigmore, *The Principles of Judicial Proof or the Process of Proof as Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials. (Second edition 1931.)*, Little, Brown and Company, Boston (Massachusetts), 1913.
[39] J. H. Wigmore, *The Principles of Judicial Proof or the Process of Proof as Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials, 2nd ed*, Little, Brown and Company, Boston (Massachusetts), 1931.