Value-aligning legitimate robot judges using λ-calculus

Evan Iatrou and Rui Li

[1] Institute of Logic Language and Computation,
University of Amsterdam, Amsterdam, The Netherlands
evangelos.iat@gmail.com
[2] Université Paris Cité, Paris, France
rui.li1@etu.u-paris.fr

Abstract

In our presentation, we intend to showcase the central role that formal philosophy ought to play in the engineering of AI models that exercise judicial power *legitimately*. We begin by arguing that according to the value of legitimacy, such AI models ought to provide justifications for their outputs that have the same logical form as the justifications found in case-law. We further argue that the conceptual re-engineering methods of Carnapian explication & narrow reflective equilibrium can be combined to guide formal philosophers in the practice of descriptively formalising those logical forms. Finally, we exhibit how this can be done when the formal philosopher uses λ-calculus to perform such formalisations. To make our case, we refer to two state-of-the-art methodologies of logically reconstructing judicial reasoning, those of LogiKEy & Catala.

1 Introduction

Algocracies, i.e., political orders where political power is exercised inter alia *by* or *via* algorithms, are already a reality [9]. A prime example of such algorithmic political authorities are AI models that exercise judicial power. Such is the example of AI that partakes in the interpretation & application of the law, the so-called "*robot judges*"[22] (e.g., identifying irregularities in contracts between the public & private sector [10], predicting the probability of recidivating [23], deciding the outcome of a criminal law case [24]). Since such AI models exercise judicial power, they now constitute political authorities themselves, and hence, they should be checked & balanced so as to avoid any abuse of power. Such a check & balance is the alignment of those models towards the value of legitimacy: *is the exercise of power by the robot judge legitimate* (*cf.* [9])?

The practice of engineering AI that abides by specific values is known in the literature as *value-alignment* [23]. Despite its seemingly "*practical*" character, value-alignment lies at the core of a traditional problem in diverse fields of analytic philosophy (e.g., philosophy of logic, (meta-)ethics, philosophy of science), that of deciding the truth value of the so-called *evaluative judgements*. Specifically, an evaluative judgement is a judgement of whether a particular value (e.g., the value of legitimacy) is applicable to a particular case (e.g., the exercise of power by a judge) ([18]; *cf.*[20]). In analytic meta-semantics, an evaluative judgement is essentially construed as a question of whether an object is subsumed by a concept or whether a term is subsumed by a specific predicate or whether a particular is an instantiation of a universal [2, 15]. The challenge with evaluative judgements' truth value is that evaluative judgements are accused of not being objective [18]. For instance, since the value of legitimacy has a different meaning in the European Enlightenment-rooted political tradition than the Chinese Confucian-based one[11], the evaluative judgement of whether a judicial authority is legitimate has a different answer in each of the two traditions. A method to overcome this (seemingly) subjective nature of evaluative judgement is to reduce it to specific *factual* judgements, where the latter can be minimally construed as empirical judgments about states of the physical world. I.e., judgements

about a *non*-subjective *ordo essendi* for which there is a strong intersubjective consensus about its characteristics. E.g., a judge's decision is legitimate only if the *factual* condition of the judge not being bribed is true. In practice, before an evaluative judgement is (partially) reduced to factual judgements, it is customary to reduce it to other evaluative judgements and then reduce those evaluative judgements to factual ones (*see e.g.* [1, 14]). Such reductions of evaluative judgement to factual ones are called *operationalisations* [7]. In our case, the evaluative judgment of AI's legitimacy-alignment should be operationalised minimally to AI's alignment towards the following two values: (i) the epistemic value of foreseeability. Specifically, the application of the law should be foreseeable.; (ii) the legal value of legality which can be minimally construed as the so-called supremacy of law: everyone, even those that exercise power, should abide by the law [8, 14].

The necessity of alignment towards the values of legality & foreseeability imposes to the robot judge the factual requirement of providing a justification for its output that has the *same* logical form with the justifications provided by the judicial authorities that have the authority by the law to judge the cases in question. In particular, for the application of the law to be foreseeable, AI should apply the *same* reasoning methods in *similar* circumstances. Otherwise, one can not know with certainty the circumstances under which they violate the law. However, different judicial authorities propose the application of different reaosning methods to similar cases raising the question of which one should the robot judge adopt. For instance, in order to establish a causal relation between an action of a defendant (e.g., shooting the victim) and the alleged harm induced by that action (e.g., the victim dying), Anlgo-American criminal law courts often employ the so-called *but-for test*: but-for the action of the defendant (legal cause), the harm (legal effect) would not have happened [17]. However, the European Court of Human Rights (ECtHR) has explicitly rejected the but-for test as a reasoning method of establishing legal causal relations between actions and harm (*E. and others v. UK*, no. 33218/96, 2002, ¶99; *cf.* [21]). Subsequently, different construals of causal reasoning can lead to different (conflicting) judgements making the application of the law less foreseeable.

The foregoing bring about the question of which should be the reasoning methods used in the application of the law. E.g., should we use the but-for test or not? The answer is given by the value of legality. Specifically, the supremacy of law dictates that the legitimate reasoning methods are those employed by the judicial authorities that are prescribed by law to judge the case in question (e.g., criminal courts judging criminal law cases). Any divergence from the reasoning methods found in the case law of those authorities undermines the value of legality.

Considering the above, we ought to impose specific logical constraints to AI that exercises judicial power so as to output justifications of the desired logical form (for an overview of such logical constraints to connectionist AI *see* [12]). Therefore, one needs to logically reconstruct the judicial reasoning used by the legitimate judicial authorities so as to determine which should those logical constraints be. Which then brings about the question of which *methodology* should one use to perform such descriptive logical reconstructions of judicial reasoning.

Two such methodologies of logical reconstruction are LogiKEy[1] and Catala,[2] which can be abstracted in the following four-steps high-level schema:[3] (i) parsing case & statutory law documents so as to mine in natural language the reasoning methods a judicial authority uses; (ii) choosing an *object logic* to formally model the mined reasoning methods. In Catala, they choose a prioritised default typed logic. The priority relation is used to model exceptions in the application of the law by prioritising conditions that override the default way that the law

[1] Introduced in [4]. The acronym stands for **Logi**c and **K**nowledge **E**ngineering Framework and Methodolog**y**.
[2] Introduced in [16]. It is "*[n]amed after Pierre Catala [...] a pioneer of French legal informatics*"[13].
[3] Note that there can (and should) be a fluctuation among those steps, albeit the basic order remains as is.

is applied. In juxtaposition to the choice of a unique object logic, LogiKEy can be employed for (a combination of) different object logics like modal logics of preferences or dyadic deontic logics [3, 4].; (iii) *translating* the logical reconstructions of the object logic to λ-terms in a target λ-calculus. LogiKEy λ-translates formulae ϕ_i of the object logic while Catala λ-translates terms t_i.; (iv) translating those λ-translations to (functional) programming code. In Catala, this code is directly used to engineer AI, while in LogiKEy, it is used by automated theorem provers (ATPs) for verifying whether the object logic is a faithful reconstruction of judicial reasoning (more on *faithfulness* below). Catala can generate code in programming languages from diverse programming paradigms (e.g., OCaml, (Java)Script, Python), while LogiKEy can employ only programming languages used by ATPs that use HOL (e.g., Isabelle/HOL).

The existence of different methodologies of logically reconstructing judicial reasoning like Catala & LogiKEy necessitates the employment of a *methodology* that can evaluate which of the available logical reconstruction methodologies is more adequate. Since a *good* methodology for logically reconstructing judicial reasoning is a methodology that generates *good* models of that reasoning, the evaluation of that methodology's goodness is reduced to the evaluation of goodness of the models that it generates. To evaluate a model's goodness, we can identify a list of criteria that a good model should satisfy. For such a list, we have to look no further than formal philosophy's early pioneer Rudolf Carnap's *explication* criteria. Specifically, explication is a conceptual re-engineering method of identifying a particular concept in a particular discourse (e.g., the concept of causal reasoning in the discourse of the ECtHR), and then, re-engineering this concept in a (non-)formal form in a different discourse (e.g., the discourse of formal philosophy of law) such that the re-engineered concept being *similar* to the initial concept, more *exact*, more *fruitful*, and more *simple*. I.e., the re-engineered concept needs to corroborate the epistemic values of similarity, exactness, fruitfulness, and simplicity. The initial concept is called *explicandum* and its explicated form is called *explicatum*. In our case, the reasoning method used by judicial authorities (e.g., causal reasoning) is the *explicandum* and its model in the object logic is the *explicatum*. The λ-translations of the object logic and the subsequent (functional) programming codes are *not* explicata since they are formalisations of the object logic in the *same* discourse as the object logic.

Considering the above, one should prefer the logical reconstruction methodologies that corroborate explication's four epistemic values the strongest. Subsequently, the choice of a logical reconstruction methodology becomes a *value*-driven decision. Hence, we are faced once more with the problem of the objectivity of evaluative judgements. We can once more overcome it by operationalising the four evaluative judgements. Since we want to use those operationalisations to engineer legitimate models of judicial reasoning, they should be grounded on the operationalisation of the value of legitimacy (e.g., choose operational definitions of similarity that make the application of the law more foreseeable). In what follows, we briefly exhibit how such an operationalisation of similarity can be used to compare LogiKEy's & Catala's adequacy.[4] We focus on the role that λ-calculus plays in the two methodologies due to its key contribution to the corroboration of the value of *similarity*.

The first *similarity* operational requirement is imposed by the value of foreseeability: legitimacy-aligned logical reconstruction need to make explicit necessary and/or sufficient conditions under which a reasoning method should be used to interpret the law so as for that interpretation to be foreseeable. I.e., the *explicandum* & the *explicatum* need to be *intentionally* similar. Intentional similarity conditions are many times determined by specific theories of interpretation of the law. In LogiKEy, λ-translations can be used to model such interpretation theories in ATPs. Through those ATP embeddings, we can evaluate whether those theorems

[4] A more thorough operationalisation of the 4 values can be found in one of the authors' MSc Thesis: [14].

are true in the object logic. The use of HOL in the LogiKEy approach is necessary for the theorem verification task. Specifically, in order to evaluate the truth of theorems expressed in λ-terms, one has to provide a semantical interpretation of those terms that allows those theorems to be truth bearers. To deal with this challenge, LogiKEy's λ-translations in HOL have bodies with the same syntactic structure as formulae of classical higher-order logic (e.g., $\lambda w.\varphi(w) \wedge \psi(w)$). *Via* this syntactical similarity, λ-translations can be semantically interpreted using Henkin models. Thus, if one uses a Henkin-sound ATP, then a proof of a theorem entails its truth. The only thing left is to prove that there is a truth preservation between the HOL's & object logic's semantics, what in [4] call as *faithfulness of the embedding*. *Contra* to LogiKEy, Catala's λ-translations do not have a semantical interpretation, and hence, its current form does not suffice to verify the truth of interpretation theories.

However, both Catala and LogiKEy can be used to corroborate the similarity requirement of *coherence*: a formal model of a judicial reasoning should verify paradigmatic applications of that reasoning method in judicial judgements so as to corroborate the coherence among those judgements. The requirement for coherence is imposed by both the value of legality and value of foreseeability. Firstly, formal models of judicial reasoning should produce the same judgements as those of the legitimate human judicial authorities (legality). Secondly, paradigmatic cases of how a reasoning method is applied in past cases are used to determine future applications of the law (foreseeability). In philosophy of law, the construal of the epistemic value of coherence as the satisfaction of paradigmatic judgements is *on par* with another landmark conceptual re-engineering method, that of *narrow reflective equilibrium* (*NRE*) advocated by pioneers in philosophy of law like John Rawls & Richard Dworkin (*see e.g.* [19]). The combination of explication and NRE can be used to balance out their perils and maximise their efficiency[6].

In order to verify paradigmatic applications of a reasoning method, both LogiKEy and Catala make use of λ-calculi's translatability to functional programming code: one can give as input to the code the facts of past cases and then verify whether the output of the code coincides with the respective judgements. For that to be possible, both methods need to employ some kind of judgement-preservation theorems: judgements that can be derived in the object logic should be derivable in the target λ-calculi as well. In the case, of LogiKEy, this is secured once more through HOL's semantical interpretation *via* Henkin models. On the other hand, in order to ensure judgement-preservation without semantics, Catala adopts the GOFAI rule-based modelling of judicial reasoning. Specifically, certain λ-translations have the syntactic structure of if-then *rules with exceptions*: the *heads* of the rules are potential judgements and the *bodies* are sufficient conditions for each of those judgements as well as possible exceptions to those conditions. By following the reduction rules of the target λ-calculus, those rule-like λ-terms can be reduced to either their heads (i.e., specific judgements) or to their exceptions. Using this rule-reduction schema, Catala ensures judgement-preservation though what they call *correctness theorem*: a rule-like λ-translation is reduced to the same λ-term that the λ-translation of the object logic's true judgement (or its true exception) is reduced to.

Summing up, it seems that LogiKEy corroborates the value of similarity better than Catala since it can be used to enhance both intentional similarity & coherence. Having said that, before concluding on which of the two methodologies is more adequate, one has to evaluate their performance in other similarity requirements (e.g., extensional & relational similarity) as well as to the rest three explication values [14, 6, 5]. Another conclusion from the above analysis is the ways that λ-calculus can be used corroborate the value of similarity setting a precedent for other logical reconstruction methodologies to follow. The foregoing remarks are a quick taste of how explication & NRE can guide a formal philosophers practice of modelling *legitimate* models of judicial reasoning that can be employed by robot judges.

References

[1] Evgeni Aizenberg and Jeroen van den Hoven. Designing for human rights in AI. *Big Data & Society*, 7(2):2053951720949566, 2020.

[2] Carlos E. Alchourrón. *Limits of logic and legal reasoning*. Reprint edition, 2015.

[3] Christoph Benzmüller, David Fuenmayor, and Bertram Lomfeld. Modelling value-oriented legal reasoning in LogiKEy. *Logics*, 2(1):31–78, 2024.

[4] Christoph Benzmüller, Xavier Parent, and Leendert van der Torre. Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial Intelligence*, 287:103348, 2020.

[5] Georg Brun. Explication as a method of conceptual re-engineering. *Erkenntnis*, 81:1211–1241, 2016.

[6] Georg Brun. Conceptual re-engineering: from explication to reflective equilibrium. *Synthese*, 197:925–954, 2020.

[7] Hasok Chang. Operationalism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[8] Venice Commission. *Rule of law checklist (CDL-AD(2016)007)*. Venice, 11-12 March 2016.

[9] John Danaher. The threat of algocracy: Reality, resistance and accommodation. *Philosophy and Technology*, 29(3):245–268, 2016.

[10] Natasha Donn. *Artificial intelligence can identify risks in public contracting court*. Portugal Resident, 03 2023.

[11] Pascale Fung and Hubert Etienne. Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between china and the EU. *AI and Ethics*, 2022.

[12] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In Lud De Raedt, editor, *IJCAI-22*, pages 5478–5485, 7 2022.

[13] Liane Huttner and Denis Merigoux. Catala: Moving towards the future of legal expert systems. *Artificial Intelligence and Law*, 2022.

[14] Evan Iatrou. Reclaiming enlightenment: On the logical foudnations of the rule of law in a legitimate algocracy. Master's thesis, Institute of Logic, Language and Computation, Univerity of Amsterdam, 10 2023.

[15] Neil MacCormick. Legal deduction, legal predicates and expert systems. *International Journal for the Semiotics of Law*, 5:181–202, 1992.

[16] Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. Catala: a programming language for the law. *Proceedings of the ACM on Programming Languages*, 5(ICFP):1–29, 2021.

[17] Michael S. Moore. Causation in the law. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019.

[18] Hilary Putnam. *The collapse of the fact/value dichotomy and other essays*. Harvard University Press, 2002.

[19] François Schroeter, Laura Schroeter, and Kevin Toh. A new interpretivist metasemantics for fundamental legal disagreements. *Legal Theory*, 26(1):62–99, 2020.

[20] Amartya K. Sen. The nature and classes of prescriptive judgments. *The Philosophical Quarterly*, 17(66):46–62, 1967.

[21] Gemma Turton. Causation and risk in negligence and human rights law. *The Cambridge Law Journal*, 79(1):148–176, 2020.

[22] Jasper Ulenaers. The impact of artificial intelligence on the right to a fair trial: Towards a robot judge? *Asian Journal of Law and Economics*, 11(2):20200008, 2020.

[23] Christoph Winter, Nicholas Hollman, and David Manheim. Value alignment for advanced artificial judicial intelligence. *American Philosophical Quarterly*, 60(2):187–203, 2023.

[24] Alena Zhabina. *How China's AI is automating the legal system*. Deutsche Welle (DW), 2023.