Four Freedoms for Deontic Logic: A Framework for Scalable AI Ethics

Ali Farjami

One of the biggest challenges in applying logic to AI systems for societal and normative purposes is scalability, both in terms of design and computation. As AI grows more powerful, we need systems that not only perform well computationally but also remain adaptable, ethical, and aligned with human values. Inspired by ideas like Richard Stallman's focus on software freedom[1] and Rich Sutton's emphasis on scalable, computation-driven methods,[2] my research aims to create flexible and scalable frameworks for normative reasoning that can meet these demands in dynamic, real-world contexts.

During my Ph.D., I contributed to the LogiKEy project [1], which focuses on designing and engineering ethical reasoners, normative theories, and deontic logics. The core methodology involves semantical embeddings of deontic logics into expressive classical higher-order logic (HOL), enabling the use of off-the-shelf theorem provers and model finders for experimentation with various logics and ethico-legal domain theories.

This approach allows for the formalization and automation of complex normative reasoning within theorem provers like Isabelle/HOL. However, I recognized limitations in modeling normative systems, particularly concerning design scalability. The LogiKEy methodology primarily depends on existing logical systems in the literature. I envisioned extending this methodology to provide users the freedom to design their own normative systems tailored to their specific applications.

To make the LogiKEy methodology more customizable, I proposed the Four Freedoms for Deontic Logic framework, which I presented at the ENIGMA conference.[3] This framework is based on input/output logic, introduced by David Makinson and Leon van der Torre [2]. While they did not mathematically define the framework, they provided illustrative examples. Unlike frameworks based on metaphysical assumptions, input/output logic focuses on normative patterns, making it adaptable to various contexts [3]. I have compared the input/output logic framework to Richard Stallman's principles of free software. The first two freedoms relate to openness and transparency, while the last two emphasize distribution and adaptability for normative systems. This framework emphasizes:

- **Freedom to Choose the Logical Base of Normative Systems:** The choice of a base logic plays a crucial role in building normative systems, as different logical frameworks offer distinct advantages depending on the context and application. For example, the use of classical logic in Mimamsa deontic logic [4] aligns with the acceptance of reductio ad absurdum for resolving contradictions, while the intuitionistic logic in Talmudic deontic logic [5] emphasizes conflict resolution through external mechanisms. These examples illustrate how selecting an appropriate base logic allows normative systems to reflect the underlying principles and reasoning structures of the domains they model, highlighting the need for flexibility in foundational choices to ensure adaptability across diverse applications. The input/output logics investigated in the literature are built on top of classical propositional logic [2] and intuitionist propositional logic [6]. In my Ph.D. thesis, it has been shown that we can build the input/output version of any abstract logic [7]. We further developed the abstract algebraic logic approach to input/output logic, where the family of selfextensional logics was proposed as a general background environment for input/output logics [8]. We introduced the generalizations of several types of permission (negative, dual negative, static, dynamic), as well as their interactions with normative systems, to various families of selfextensional logics, thereby proposing a systematic approach to the definition of normative and permission systems on nonclassical propositional bases [8].

- **Freedom to Characterize Normative Systems:** In modal logic, there is some degree of freedom for characterizing normative systems through the addition or removal of axioms. However, input/output logic provides even greater flexibility, as it allows the modification of fundamental inference rules to better suit specific normative applications. This freedom is essential for tailoring normative systems to specific applications.

[1] See: Richard Stallman, "Four Freedoms"
[2] See: The Bitter Lesson
[3] See my talk at my Homepage

For instance, in discursive reasoning, modifying input/output logic by omitting the conjunction rule (AND) allows for more nuanced discourse analysis [7]. Similarly, Alexander Bochman's adaptation of input/output logic (by adding the bottom axiom) for causal reasoning demonstrates how altering logical structures can effectively model causality [9]. However, transitioning between these customized logical systems can be complex. An algebraic characterization offers a granular approach to designing and understanding normative systems, facilitating smoother transitions and more precise modeling across various logical frameworks. We further developed the algebraic approach to input/output logic initiated in [10], where subordination algebras and a family of their generalizations were proposed as a semantic framework for various input/output logics. In particular, we explored precontact algebras as a suitable algebraic environment for modeling negative permission and characterized the properties of several types of permission (negative, static, dynamic), as well as their interactions with normative systems, using appropriate modal languages to encode outputs [11].

- **Freedom to Combine Normative Systems:** Compositionality is a crucial principle in logic and normative systems, as it ensures that complex structures can be understood and constructed by combining simpler components in a systematic way. Normative systems are complex, and designing normative applications requires combining different normative components with distinct properties. For instance, various types of permissions are defined based on how permissive norms interact with obligatory norms [12]. Furthermore, constitutive norms and regulative norms must be combined to represent social phenomena effectively [13]. Thus, finding a methodological approach to characterize the possibilities for combining normative systems provides significant flexibility in their design. In particular, we have studied a set of first-order formulas, known as Kracht formulas [14], which offer a framework for integrating obligation, permission, and prohibition systems for diverse applications. We characterized the syntactic shape of first-order conditions on algebras endowed with subordination, precontact, and dual precontact relations, ensuring that these conditions correspond to axioms in the aforementioned modal language. Additionally, we introduced algorithms for computing the first-order correspondents of modal axioms on such algebras and, conversely, for computing the modal axioms whose first-order correspondents satisfy the specified syntactic shape [14].

- **Freedom to Implement and Adapt Normative Systems:** LogiKEy methodology was successful for several well-known deontic logics, including the dyadic deontic logic by Åqvist [15], as well as the more intricate one by Carmo and Jones [16]. However, since input/output (I/O) logic employs an operational semantics, the use of shallow semantical embedding was not particularly effective. There have been some translations from I/O logic into modal logic [17], but this approach doesn't provide the whole picture. One of the primary questions during my Ph.D. was to identify suitable mathematical models for input/output logic, such as Kripke semantics or more abstract ones like Boolean algebra models. Through my research, I devised two methods for mathematically formulating input/output logic.

  - *Starting from the operational semantics of I/O logic:* we can algebraically formulate input/output operations. This involves introducing algebraic operators to manipulate and correlate inputs and outputs. During my Ph.D., I implemented algebraic input/output operations, ensuring that the embedding in HOL is both sound and faithful [18]. Regarding the computational complexity of input/output operations, focusing on relational semantics for input/output logic and its implementation appears to be a promising direction.
  - *Starting from the proof systems of I/O logic:* we can begin with derivation rules such as AND, OR, and CT to construct mathematical models. In collaboration with Alessandra Palmigiano and her Ph.D. students, we discerned that the basic input/output proof system aligns with subordination algebras [10]. It is straightforward to implement the corresponding modal algebras of obligation, negative permission, and dual-negative permission aligned with original proposed syetems. The soundness result demonstrate that the efficiency of the algebraic encoding is similar to the LogiKEy benchmark examples [19].[4] By establishing a systematic connection between input/output logic and various modal algebras, our approach will allow for more efficient and accurate use of off-the-shelf theorem provers in the development of responsible AI systems, enabling the LogiKEy framework to address a wider array of logical formalisms with greater computational efficiency.

Building on the Four Freedoms framework, I am particularly interested in extending its application to dynamic normative systems. This would enable the development of adaptive and scalable frameworks that align with evolving AI technologies, ensuring ethical consistency and robust decision-making in dynamic environments.

[4] See the GitHub link (direct implementation): https://github.com/farjami110/AlgebriacInputOutput

References

[1] 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*, 237:103348, 2020.

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

[3] David Makinson. On a fundamental problem of deontic logic. *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, pages 29–54, 1999.

[4] Björn Lellmann, Francesca Gulisano, and Agata Ciabattoni. Mīmāṃsā deontic reasoning using specificity: a proof theoretic approach. *Artificial Intelligence and Law*, 29(3):351–394, 2021.

[5] Michael Abraham, Dov M Gabbay, and Uri Schild. Obligations and prohibitions in talmudic deontic logic. *Artificial Intelligence and Law*, 19:117–148, 2011.

[6] Xavier Parent, Dov Gabbay, and Leendert van der Torre. Intuitionistic basis for input/output logic. In *David Makinson on Classical Methods for Non-Classical Problems*, pages 263–286. Springer, 2014.

[7] Ali Farjami. *Discursive Input/Output Logic: Deontic Modals, and Computation*. PhD thesis, University of Luxembourg, Luxembourg, 2020.

[8] Andrea De Domenico, Ali Farjami, Krishna Manoorkar, Alessandra Palmigiano, Mattia Panettiere, and Xiaolong Wang. Obligations and permissions on selfextensional logics, 2024.

[9] Alexander Bochman. *A logical theory of causality*. MIT Press, 2021.

[10] Andrea De Domenico, Ali Farjami, Krishna Manoorkar, Alessandra Palmigiano, Mattia Panettiere, and Xiaolong Wang. Subordination algebras as semantic environment of input/output logic. In *International Workshop on Logic, Language, Information, and Computation*, pages 326–343. Springer, 2022.

[11] Andrea De Domenico, Ali Farjami, Krishna Manoorkar, Alessandra Palmigiano, Mattia Panettiere, and Xiaolong Wang. Obligations and permissions, algebraically. *arXiv preprint arXiv:2403.03148*, 2024.

[12] David Makinson and Leendert van der Torre. Permission from an input/output perspective. *Journal of Philosophical Logic*, 32(4):391–416, 2003.

[13] Xin Sun and Leendert van der Torre. Combining constitutive and regulative norms in input/output logic. In Fabrizio Cariani, Davide Grossi, Joke Meheus, and Xavier Parent, editors, *Deontic Logic and Normative Systems — 12th International Conference, DEON 2014, Ghent, Belgium, July 12-15, 2014*, pages 241–257. Springer, 2014.

[14] Andrea De Domenico, Ali Farjami, Krishna Manoorkar, Alessandra Palmigiano, Mattia Panettiere, and Xiaolong Wang. Correspondence and inverse correspondence for input/output logic and region-based theories of space. *arXiv preprint arXiv:2412.01722*, 2024.

[15] Christoph Benzmüller, Ali Farjami, and Xavier Parent. Åqvist's dyadic deontic logic E in HOL. *Journal of Applied Logics – IfCoLoG Journal of Logics and their Applications (Special Issue on Reasoning for Legal AI)*, 6(5):733–755, 2019.

[16] Christoph Benzmüller, Ali Farjami, and Xavier Parent. A dyadic deontic logic in HOL. In Jan Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems — 14th International Conference, DEON 2018, Utrecht, The Netherlands, 3-6 July, 2018*, pages 33–50. College Publications, 2018.

[17] Christoph Benzmüller, Ali Farjami, Paul Meder, and Xavier Parent. I/O logic in HOL. *Journal of Applied Logics – IfCoLoG Journal of Logics and their Applications (Special Issue on Reasoning for Legal AI)*, 6(5):715–732, 2019.

[18] Ali Farjami. New algebraic normative theories for ethical and legal reasoning in the LogiKEy framework. *arXiv preprint arXiv:2107.11838*, 2021.

[19] Christoph Benzmüller, Ali Farjami, David Fuenmayor, Paul Meder, Xavier Parent, Alexander Steen, Leendert van der Torre, and Valeria Zahoransky. LogiKEy workbench: Deontic logics, logic combinations and expressive ethical and legal reasoning (Isabelle/HOL dataset). *Data in Brief*, 33:106409, 2020.