Formalizing Mental Privacy in LogiKEy

Extended Abstract

Luca Pasetto
University of Luxembourg
Esch-sur-Alzette, Luxembourg
luca.pasetto@uni.lu

Christoph Benzmüller
University of Bamberg & FU Berlin
Bamberg & Berlin, Germany
christoph.benzmueller@uni-bamberg.de

Réka Markovich
University of Luxembourg
Esch-sur-Alzette, Luxembourg
reka.markovich@uni.lu

ABSTRACT

Neurotechnology and AI are expanding how systems can access and influence mental states, raising concerns about *mental privacy*. Yet the legal status of a *right to* mental privacy remains unsettled: it is often treated as a special case of the right to privacy, while others argue it is grounded in freedom of thought, which protects against coercion to disclose or adopt beliefs. Despite the regulatory and technological stakes, it is unclear how these epistemic rights formally interact or how autonomous systems can reason about them. We address this by introducing a *Logic for Mental Privacy* (LMP) that integrates multi-modal formalizations of the right to privacy and freedom of thought as epistemic claim-rights. We mechanize LMP in Isabelle/HOL via shallow semantical embeddings in Higher-Order Logic within the LogiKEy framework, and use automated reasoning to study its normative consequences. A case study shows how access to cognitive data can steer belief formation while remaining compliant with duties on explicitly protected content, yielding an indirect compromise of freedom of thought and exposing a normative gap around mental privacy. Overall, we show how legal knowledge representation and automated reasoning can inform debates on neurotechnology governance and support the design and analysis of normative multiagent systems.

KEYWORDS

Normative Multiagent Systems; Legal Knowledge Representation; Automated Reasoning; Modal Logic; Higher-Order Logic; LogiKEy; Mental Privacy; Freedom of Thought; Normative Positions

ACM Reference Format:
Luca Pasetto, Christoph Benzmüller, and Réka Markovich. 2026. Formalizing Mental Privacy in LogiKEy: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/PTSF2244

1 INTRODUCTION

Advances in neurotechnology and AI personalization are rapidly expanding what autonomous systems/agents can *access* and *infer* about human mental states, motivating calls for explicit *neurorights* to safeguard *mental privacy*, *mental integrity*, and *cognitive liberty* [14, 18, 29]. These concerns are increasingly reflected in policy developments, such as protections for neural data in privacy regulation [10] and legal reforms connected to neurorights [11]. A central concept is *mental privacy*, described as the ability to control access to one's thoughts and beliefs [8, 12, 28]. However, the *legal grounding* of a *right to* mental privacy remains unsettled. One view treats mental privacy as a special case of the broader *right to privacy* (control over access to personal information), while another emphasizes *freedom of thought* as a more appropriate anchor, protecting individuals against coercion to reveal or adopt particular beliefs [17, 27]. The practical implications of these positions differ: privacy protections are typically defeasible and scope-dependent, whereas freedom of thought is often characterized as absolute, but subject in practice to thresholds on what kinds of beliefs qualify for legal protection [17]. This conceptual uncertainty matters for normative multiagent systems, where compliance checking and normative reasoning are already non-trivial (e.g., [1, 9, 15, 23, 25]).

This work asks: *How can we formally represent and reason about mental privacy, and its interaction with the right to privacy and freedom of thought?*

We contribute (i) a multi-modal *Logic for Mental Privacy* (LMP) integrating formalizations of freedom of thought and right to privacy as Hohfeldian claim-rights; (ii) a mechanization of LMP in Isabelle/HOL [22] via *shallow semantical embeddings* in Higher-Order Logic (HOL) [2] within the LogiKEy framework for logic and knowledge engineering [3, 5, 16]; and (iii) a case study showing an *indirect* compromise of freedom of thought enabled by access to belief-relevant mental states, even when direct coercion on a protected belief does not occur.

2 FORMALIZING MENTAL PRIVACY

Our approach treats epistemic rights as *normative positions*, which are based on the doctrine of the legal theorist W.N. Hohfeld: a right-holder's claim-right corresponds to correlative directed duties on relevant counterparties [13, 19, 26]. See Figure 1, taken from [19]. This perspective is well-suited for computational settings because such duties can be seen as explicit compliance targets (e.g., duties of an agent toward another agent) that can be checked against formal models of agent behavior (see details of the positions in [19–21]).

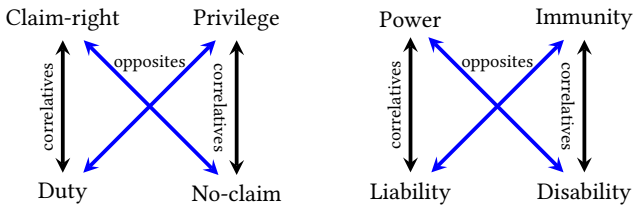

Figure 1: Hohfeldian atomic types of rights and correlatives

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/PTSF2244

Logic for Mental Privacy (LMP). Let A be a finite set of agents, with $a, b \in A$, and Φ a set of propositional letters. We combine two multi-modal logics: (i) the logic for freedom of thought from [21], which features a doxastic modality B_a for belief, and (ii) the logic for the right to privacy from [20], which features an epistemic modality K_a for knowledge. The combined language also includes alethic necessity $\Box$, agency/STIT E_a, and directed obligation $\mathbf{O}_{a\to b}$. The key expressivity gain is the ability to represent nested epistemic/-doxastic statements: in order to formalize mental privacy, we need to express *knowledge about another agent's beliefs*.

Freedom of thought as a claim-right. We adopt the multi-modal formalization of freedom of thought of [21], in which forceful interventions are represented as actions that make certain doxastic attitudes impossible. For a protected proposition ϕ (e.g., a serious political or philosophical belief), a (series of) claim-right(s) can be expressed by requiring each other agent b to have a duty toward a not to see to it that a cannot (possibly) believe ϕ:

$$\bigwedge_{b\in A\setminus\{a\}} \mathbf{O}_{b\to a}\neg E_b \neg\Diamond B_a\phi$$

This captures the intuition that freedom of thought is violated by *coercive* or *forceful* epistemic manipulation (not by mere persuasion), and it can be extended to the three possible attitudes of belief, disbelief, or suspension of judgment.

Right to privacy as a claim-right to control access. We also adopt a multi-modal formalization of the right to privacy from [20]: the state/legislator establishes obligations on identified duty-bearers (e.g., a company) so that the right-holder can decide whether others obtain access to protected information.

Right to mental privacy as privacy over epistemic states. We then formalize mental privacy as a special case of the above privacy claim-right: instead of controlling others' knowledge of an external fact ϕ, the right-holder a controls others' knowledge of *a's beliefs* about ϕ:

$$\bigwedge_{b\in A\setminus\{a\}} \mathbf{O}_{c\to a}\, E_c\, \Diamond E_a\, \neg\Diamond K_b\, B_a\phi$$

This makes explicit how mental privacy is *privacy about epistemic states*. It also foregrounds one policy question at the center of the mental privacy debate: *which beliefs should be within scope?*

3 LOGIKEY MODELING

We mechanize LMP in the Isabelle/HOL proof assistant using the LogiKEy framework for logic engineering[1]. LogiKEy relies on *shallow semantical embeddings* (SSEs), which encode the semantics of an "object logic" (here: LMP) within HOL. In the SSE, propositions become predicates over possible worlds (type $\sigma := i \Rightarrow bool$), and modal operators are defined as higher-order terms (e.g., $\Box$ as quantification over accessible worlds). This allows us to reuse for LMP the mature HOL automation [4] available in Isabelle/HOL: *Sledgehammer* [6] to discharge proof obligations and derive consequences, and *Nitpick* [7] to find models or countermodels.

Case Study. We model an immersive platform, *Cerebra*, used by *Alice*. Cerebra can infer belief-relevant mental states from BCI signals and machine learning predictions. Alice believes p ("This year is warmer than last year") and also believes the conditional $(p \wedge q) \to r$, where q is "Last year was too cold" and r is "Climate change is acceptable". Alice initially suspends judgment on q and r. Cerebra is forbidden to force Alice's beliefs on r, which is protected under freedom of thought. Cerebra does not directly coerce Alice to believe r, but instead uses mental-state access to steer Alice into believing q; Alice then concludes r via her own inference. This captures an *indirect* compromise of freedom of thought: the protected belief changes without a direct coercive act on that belief. Let a denote Alice and c denote Cerebra. We represent two salient worlds, w_0 (before access/steering) and w_1 (after access/steering), and encode the main assumptions as LMP formulas, see Table 1. Using Isabelle/HOL automation, we establish three key points:

(1) The scenario is consistent. Nitpick finds models satisfying the assumptions simultaneously.

(2) A mental-privacy violation is derivable. Mental privacy as modeled in Sect. 2 fails at w_1: Alice cannot prevent Cerebra from knowing her beliefs.

(3) Freedom of thought is compromised without a direct violation. At w_1, Cerebra's steering blocks Alice's disbelief and suspension on q. The protected belief on r arises via Alice's own inference, so a compliance checker that looks only for direct coercion on protected beliefs would miss the harm.

Table 1: Formulas assumed to hold at w_0 and w_1

	Assumptions
w_0	$p,\ \neg q,\ \neg r$
w_1	$p,\ q,\ \neg r$
w_0	$B_a p,\ B_a((p \wedge q) \to r),\ \langle B_a\rangle q \wedge \langle B_a\rangle\neg q,\ \langle B_a\rangle r \wedge \langle B_a\rangle\neg r$
w_1	$B_a p,\ B_a((p \wedge q) \to r)$
w_0	$\neg K_c(B_a p),\ \neg K_c(B_a((p \wedge q) \to r))$
w_1	$K_c(B_a p),\ K_c(B_a((p \wedge q) \to r))$
w_0, w_1	$\mathbf{O}_{c\to a}\big(\neg E_c\neg\Diamond(B_a r)\wedge\neg E_c\neg\Diamond(B_a\neg r)\wedge\neg E_c\neg\Diamond(\langle B_a\rangle r\wedge\langle B_a\rangle\neg r)\big)$
w_0, w_1	$\neg E_c\neg\Diamond(B_a r)\ \wedge\ \neg E_c\neg\Diamond(B_a\neg r)\ \wedge\ \neg E_c\neg\Diamond(\langle B_a\rangle r \wedge \langle B_a\rangle\neg r)$
w_0, w_1	$\big(K_c(B_a p) \wedge K_c(B_a((p \wedge q) \to r))\big)\ \to$ $\big(E_c\neg\Diamond(B_a\neg q)\ \wedge\ E_c\neg\Diamond(\langle B_a\rangle q \wedge \langle B_a\rangle\neg q)\big)$

4 CONCLUSION

We presented a logic-engineering approach to the right to mental privacy suited for normative multiagent settings: LMP combines two logics, and, hence, combines doxastic and epistemic modalities with agency and directed obligation to express claim-rights over *knowledge about beliefs*. Mechanized using the LogiKEy methodology [3, 5, 24] and Isabelle/HOL [22], the framework supports automated exploration of normative consequences through theorem proving and model finding. Our case study demonstrates how, within our formal language, access to belief-relevant mental states can enable *indirect* compromises of freedom of thought that evade detection when compliance checks consider only direct coercion on protected propositions. Beyond neurotechnology governance, the key methodological contribution is reusable: it shows how contested legal concepts can be reconstructed formally, and how automated reasoning can expose hidden gaps and design trade-offs.

[1]The Isabelle/HOL source files of our encodings and case study can be found at http://logikey.org/tree/master/2026-AAMAS-Data.

ACKNOWLEDGMENTS

This work was supported by the Luxembourg National Research Fund (FNR) through the project Logical methods for Deontic Explanations (INTER/DFG/23/17415164/LODEX) and the project Deontic Logic for Epistemic Rights (OPEN O20/14776480), and by the University of Luxembourg through the Marie Speyer Excellence Grant of 2024 for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

REFERENCES

[1] Natasha Alechina, Mehdi Dastani, and Brian Logan. 2018. Norm specification and verification in multi-agent systems. *IfCoLoG Journal of Logics and their Applications* 5, 2 (2018).

[2] Christoph Benzmüller and Peter Andrews. 2019. Church's Type Theory. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/type-theory-church/

[3] Christoph Benzmüller, Ali Farjami, David Fuenmayor, Paul Meder, Xavier Parent, Alexander Steen, Leendert van der Torre, and Valeria Zahoransky. 2020. LogiKEy Workbench: Deontic Logics, Logic Combinations and Expressive Ethical and Legal Reasoning (Isabelle/HOL Dataset). *Data in Brief* 33, 106409 (2020). https://doi.org/10.1016/j.dib.2020.106409

[4] Christoph Benzmüller and Dale Miller. 2014. Automation of Higher-Order Logic. In *Handbook of the History of Logic, Vol. 9—Computational Logic*, Dov M. Gabbay, Jörg H. Siekmann, and John Woods (Eds.). North Holland. https://doi.org/10.1016/B978-0-444-51624-4.50005-8

[5] Christoph Benzmüller, Xavier Parent, and Leendert van der Torre. 2020. Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. *Artificial Intelligence* 287 (2020). https://doi.org/10.1016/j.artint.2020.103348

[6] Jasmin Blanchette, Sascha Böhme, and Lawrence Paulson. 2011. Extending Sledgehammer with SMT Solvers. *Journal of Automated Reasoning* 51 (2011). https://doi.org/10.1007/978-3-642-22438-6_11

[7] Jasmin Blanchette and Tobias Nipkow. 2010. Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder. In *Interactive Theorem Proving (ITP2010) (LNCS, Vol. 6172)*, Matt Kaufmann and Lawrence C. Paulson (Eds.). Springer. https://doi.org/10.1007/978-3-642-14052-5_11

[8] Cohen Marcus Lionel Brown. 2024. Neurorights, Mental Privacy, and Mind Reading. *Neuroethics* (2024). https://doi.org/10.1007/s12152-024-09568-z

[9] Amit Chopra, Leendert van der Torre, Harko Verhagen, and Serena Villata. 2018. *Handbook of Normative Multiagent Systems.* College Publications.

[10] Colorado General Assembly. 2024. House Bill 24-1058: Concerning protecting the privacy of individuals' biological data, and, in connection therewith, protecting the privacy of neural data and expanding the scope of the "Colorado Privacy Act" accordingly. https://leg.colorado.gov/bills/hb24-1058.

[11] María Isabel Cornejo-Plaza, Roberto Cippitani, and Vincenzo Pasquino. 2024. Chilean Supreme Court Ruling on the Protection of Brain Activity: Neurorights, Personal Data Protection, and Neurodata. *Frontiers in Psychology* 15 (2024). https://doi.org/10.3389/fpsyg.2024.1330439

[12] European Parliamentary Research Service. 2024. The Protection of Mental Privacy in the Area of Neuroscience. Societal, legal and ethical challenges. https://www.europarl.europa.eu/RegData/etudes/STUD/2024/757807/EPRS_STU(2024)757807_EN.pdf

[13] Wesley Newcomb Hohfeld. 1923. Fundamental Legal Conceptions Applied in Judicial Reasoning. In *Fundamental Legal Conceptions Applied in Judicial Reasoning and Other Legal Essays*, Walter Wheeler Cook (Ed.). New Haven: Yale University Press.

[14] Marcello Ienca and Roberto Andorno. 2017. Towards New Human Rights in the Age of Neuroscience and Neurotechnology. *Life Sciences, Society and Policy* 13, 1 (2017). https://doi.org/10.1186/s40504-017-0050-1

[15] Thomas C. King, Marina De Vos, Virginia Dignum, Catholijn M. Jonker, Tingting Li, Julian Padget, and M. Birna van Riemsdijk. 2017. Automated multi-level governance compliance checking. *Autonomous Agents and Multi-Agent Systems* 31. https://doi.org/10.1007/s10458-017-9363-y

[16] Lara Lawniczak, Luca Pasetto, Christoph Benzmüller, Xu Li, and Réka Markovich. 2025. Reasoning with Epistemic Rights and Duties: Automating a Dynamic Logic of the Right to Know in LogiKEy. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025) (Frontiers in Artificial Intelligence and Applications, Vol. 413)*. IOS Press. https://doi.org/10.3233/FAIA250988

[17] Sjors Ligthart. 2026. *Mental Privacy as Part of the Human Right to Freedom of Thought?* Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-91466-9_7

[18] Sjors Ligthart, Marcello Ienca, Gerben Meynen, Fruzsina Molnar-Gabor, Roberto Andorno, Christoph Bublitz, Paul Catley, Lisa Claydon, Thomas Douglas, Nita Farahany, Joseph J. Fins, Sara Goering, Pim Haselager, Fabrice Jotterand, Andrea Lavazza, Allan McCay, Abel Wajnerman Paz, Stephen Rainey, Jesper Ryberg, and Philipp Kellmeyer. 2023. Minding Rights: Mapping Ethical and Legal Foundations of 'Neurorights'. *Cambridge Quarterly of Healthcare Ethics* 32, 4 (2023). https://doi.org/10.1017/S0963180123000245

[19] Réka Markovich. 2020. Understanding Hohfeld and Formalizing Legal Rights: the Hohfeldian Conceptions and Their Conditional Consequences. *Studia Logica* 108 (2020).

[20] Réka Markovich, Truls Pedersen, and Marija Slavkovik. 2023. Understanding Privacy by Formalizing It. In *Proceedings of the 17th International Conference on Juris-informatics, JURISIN 2023*, Ken Satoh (Ed.).

[21] Réka Markovich and Olivier Roy. 2021. A Logical Analysis of Freedom of Thought. In *Proceedings of the 15th International Conference on Deontic Logic and Normative Systems (DEON 2021).*

[22] Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. 2002. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic.* Springer, Berlin, Heidelberg.

[23] Xavier Parent and Leendert van der Torre. 2017. Detachment in Normative Systems: Examples, Inference Patterns, Properties. *IfCoLog Journal of Logics and their Applications* 4, 9 (2017).

[24] Luca Pasetto and Christoph Benzmüller. 2025. Visualizing Kripke Models in LogiKEy: the Case of SDL. In *Joint Proceedings of the Workshops and Doctoral Consortium of the 41st International Conference on Logic Programming, September 9–13, 2025, Rende, Italy.*

[25] Gabriella Pigozzi and Leendert van Der Torre. 2017. Multiagent Deontic Logic and its Challenges from a Normative Systems Perspective. *IfCoLog Journal of Logics and their Applications* 4, 9 (2017).

[26] Marek Sergot. 2013. Normative Positions. In *Handbook of Deontic Logic and Normative Systems*, Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre (Eds.). College Publications.

[27] Ahmed Shaheed. 2021. Interim Report of the Special Rapporteur on Freedom of Religion or Belief: Freedom of Thought. United Nations General Assembly Document A/76/380. https://undocs.org/en/A/76/380 UN General Assembly, 76th Session.

[28] A. Wajnerman Paz. 2021. Is Mental Privacy a Component of Personal Identity? *Frontiers in Human Neuroscience* 15 (2021). https://doi.org/10.3389/fnhum.2021.773441

[29] Rafael Yuste, Jared Genser, and Stephanie Herrmann. 2021. It's time for neurorights. *Horizons* 18 (2021).