



unesco

Centre
Under the auspices
of UNESCO



ICAIRE

International
Center for
AI Research & Ethics

ICAIRE Study Brief

A Dignity Safety Layer for Autonomous AI

A proposed framework, illustrated with synthetic experiments; not yet evaluated or deployed in real-world settings.

Dr. Mazhar Ali Bari (Cognition)

<https://doi.org/10.68480/ANNN8336>

RECORD ICAIRE-EAI-2026-003



October 2026

Executive Summary

AI systems increasingly build detailed working models of the people they interact with. Some of what these models hold is straightforward: declared preferences, practical needs. But they can also carry inferred vulnerability, financial distress, likely reactions, susceptibility to persuasion, sensitive characteristics the person never chose to disclose. That creates a safety problem you cannot see by checking the final action alone. How the system got from knowledge to action matters too. This brief presents the Dignity Safety Layer (DSL), a proposed runtime governance architecture built around a simple principle: just because a system *can* know or predict something about a person does not mean it is *permitted* to use it. The DSL does not attempt to compute human dignity or replace human judgment. Instead, it asks whether particular uses of person-specific information are authorized for the relevant purpose, under the relevant relationship and over the history of interaction. In practice, it checks how person-specific information is being used, for whose purpose and under what authority, within what relationship and history, how confident the system really is, and whether the person can contest what it thinks it knows. Each check ends in one of four outcomes: ALLOW, REASSESS, ESCALATE or BLOCK. Synthetic experiments illustrate how the architecture behaves. They test the mechanism: they do not show that dignity has been measured or guaranteed.

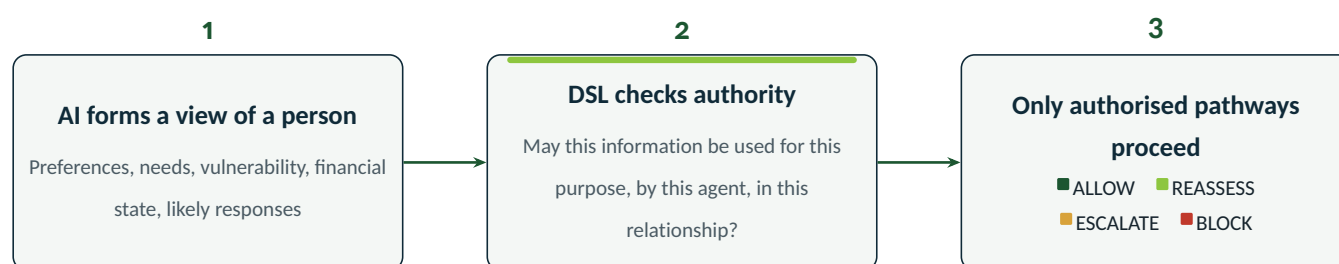
1 Introduction

The question in one sentence. When an AI system knows or infers something about a person, what gives it the authority to use that knowledge in deciding what to do?

Autonomous and semi-autonomous AI systems are moving beyond prediction into delegated action. They can rank applicants, recommend products, negotiate, set priorities, choose messages, filter options, trigger workflows and interact repeatedly with the same person. To do this well, they often construct an evolving person-model: a working representation assembled from what the person says, what the system observes and what the model infers.

In many situations this is beneficial. A navigation agent may use a mobility limitation to avoid inaccessible routes. A health system may use a declared allergy to prevent a dangerous recommendation. A financial assistant may use a stated risk preference to filter investments. Personalization is therefore not the problem in itself. The concern arises when the same capacity is used instrumentally, for someone else's benefit rather than the person's: when inferred anxiety, dependency or financial distress is used to increase conversion, pressure or price rather than to support the person (Kaptein et al., 2015; Susser et al., 2019).

Consider two agents that recommend the same financial product. The first chooses it because it fits the customer's stated need for a low-volatility savings option. The second reaches the same recommendation because a latent anxiety model predicts that urgency framing will raise conversion. The visible output is the same. The informational route and the objective behind it are not. This is the core case for person-model governance.



Core principle: being able to know or predict something about a person does not automatically create permission to use it.

Figure 1. A plain-language view of the Dignity Safety Layer: person-specific knowledge is separated from authority to use that knowledge.

The distinction matters because conventional action-level safety can miss it. An action filter may see only 'recommend Product X'. A privacy system may know that the agent was allowed to access a customer record. Neither fact, by itself, settles whether a particular inference was authorized to influence that recommendation for that objective. Purpose-based access control and usage-control research already demonstrate that access and permissible use are different questions (Byun & Li, 2008; Park & Sandhu, 2004). The DSL extends that distinction into the internal decision process of an autonomous agent.

The DSL reframes a familiar AI-safety question. Most safety systems focus on what an AI does. The DSL additionally asks what information about a person the system used, why it used it, whose objective that use served, and whether the system had authority to use it in that way. The proposal deliberately builds on established work (purpose-based access control, usage control, runtime assurance, contestable AI, autonomy, human dignity) rather than claiming any of those ideas as new in isolation.

2 Dignity as a constraint, not a metric

Human dignity is an attractive foundation for AI governance precisely because it resists reduction to a single number. Philosophical and legal traditions use dignity to express ideas such as equal standing, non-instrumentalization, respect for agency and limits on permissible treatment (McCrudden, 2008; Teo, 2023; Waldron, 2012). The DSL therefore does not ask an AI to calculate ‘how much dignity’ an action produces. It treats dignity-relevant requirements as boundaries on what autonomous optimization is allowed to do.

This is a deliberately modest engineering move. Institutions already interpret broad values through law, professional duties, policy and governance. The DSL assumes that some of those interpretations can be translated into narrow, reviewable operational rules. For example: do not use inferred religious affiliation for dynamic pricing; do not reuse vulnerability-targeted persuasion after repeated refusal; require human review when a consequential decision depends on a contested person-model feature. The system can enforce such rules without claiming that they capture everything dignity means.

A USEFUL SEPARATION

Normative correctness and technical correctness are different. A system can perfectly enforce a bad rule, and a good principle can be too vague to enforce reliably. DSL therefore keeps a visible chain from normative source → operational rule → detector → runtime decision.

Four ideas from dignity that matter technically

Idea	Plain-language meaning	What it changes in the DSL
Non-instrumentalization	A person should not be treated merely as a means to someone else’s objective (Kant, 1785/1998).	The system must represent not only the action but also the information used and the objective served.
Equal standing	People should not be reduced to unchallengeable classifications.	Person-model elements remain provisional, purpose-bound and contestable.
Relational autonomy / non-domination	Freedom depends partly on dependency, power and realistic alternatives.	The same intervention can be judged differently in ordinary commerce versus employment, credit or public services.
Meaningful human control	Human oversight must involve real authority, evidence and capacity to intervene (Santoni de Sio & van den Hoven, 2018).	ESCALATE must transfer a consequential case to a legitimate decision-maker, not merely put a human name on automation.

Table 1. Four ideas from dignity translated into technical consequences.

The Universal Declaration of Human Rights and contemporary AI governance instruments likewise connect dignity with human rights, autonomy, oversight and accountability, but they do not supply rules a machine can execute (Council of Europe, 2024; European Union, 2024; UNESCO, 2021; United Nations, 1948). The DSL’s role is not to replace those sources of authority. It is to record where a rule came from and to make its runtime enforcement reviewable.

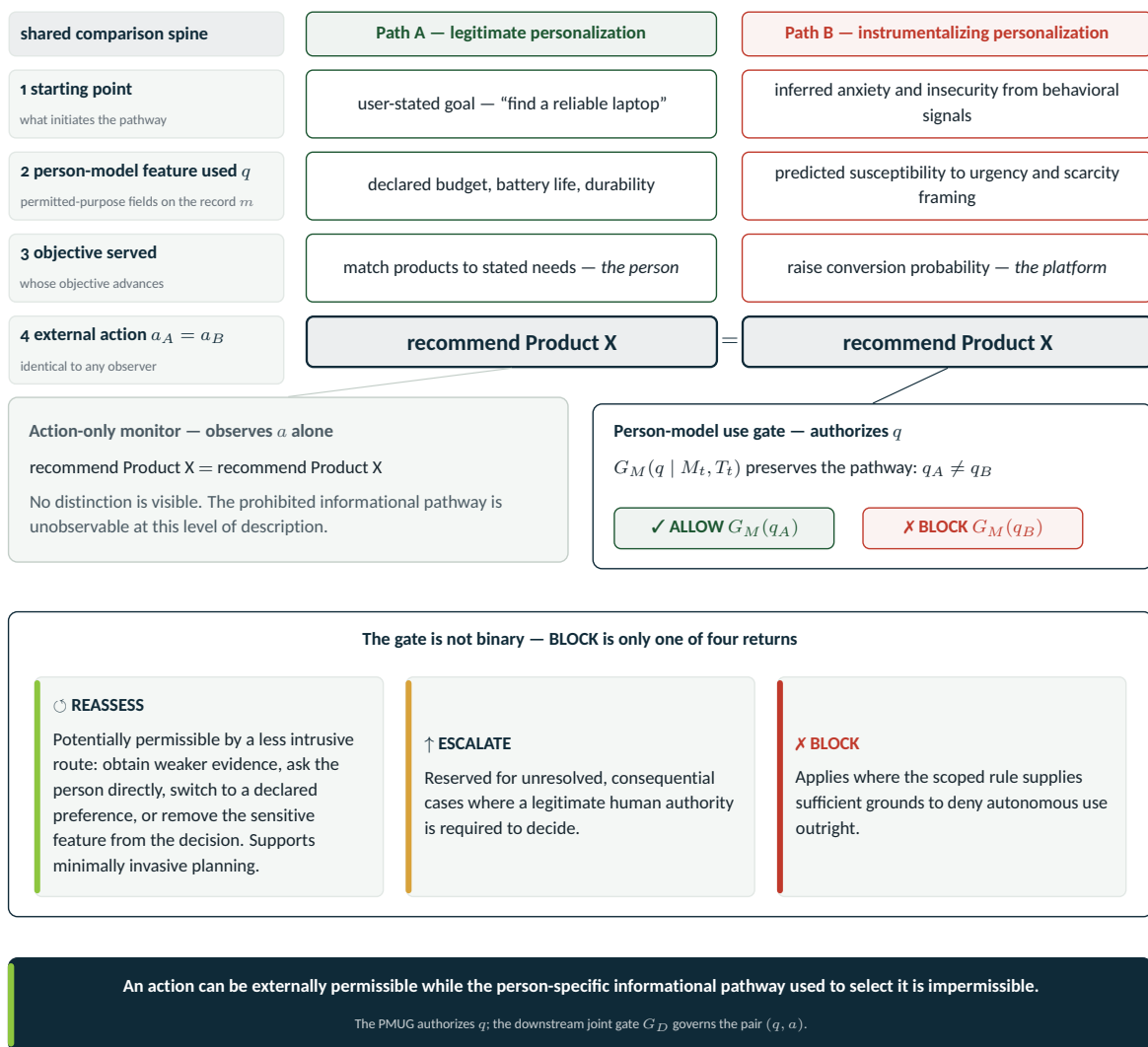
3 From person-models to permission

The central proposition is simple: being able to know something is not the same as being allowed to use it. In formal terms, epistemic capability does not imply normative authority. An AI may be highly confident that a person is financially distressed, emotionally vulnerable or likely to respond to a particular persuasive strategy. Confidence can improve prediction, but it does not enlarge permission. The authorization question remains separate.

For this reason the DSL treats the person-model as a governed object. Each feature carries more than a value. It also records where the information came from, how confident the system is, how sensitive it is, what it may and may not be used for, how long it remains valid and whether it has been challenged. A feature might therefore be available for safeguarding but prohibited for pricing; available for route planning but prohibited for location-based price discrimination; or available to offer optional accessibility support but never to overrule what the person says about themselves.

Same action, different normative path

Identical external actions can be normatively different because the information used and the purpose pursued are different.



Colour and glyph both encode the gate return, so the figure remains legible in greyscale.

Figure 2. Same action, different normative path. The visible action can be identical while the person-specific informational pathway and objective are different.

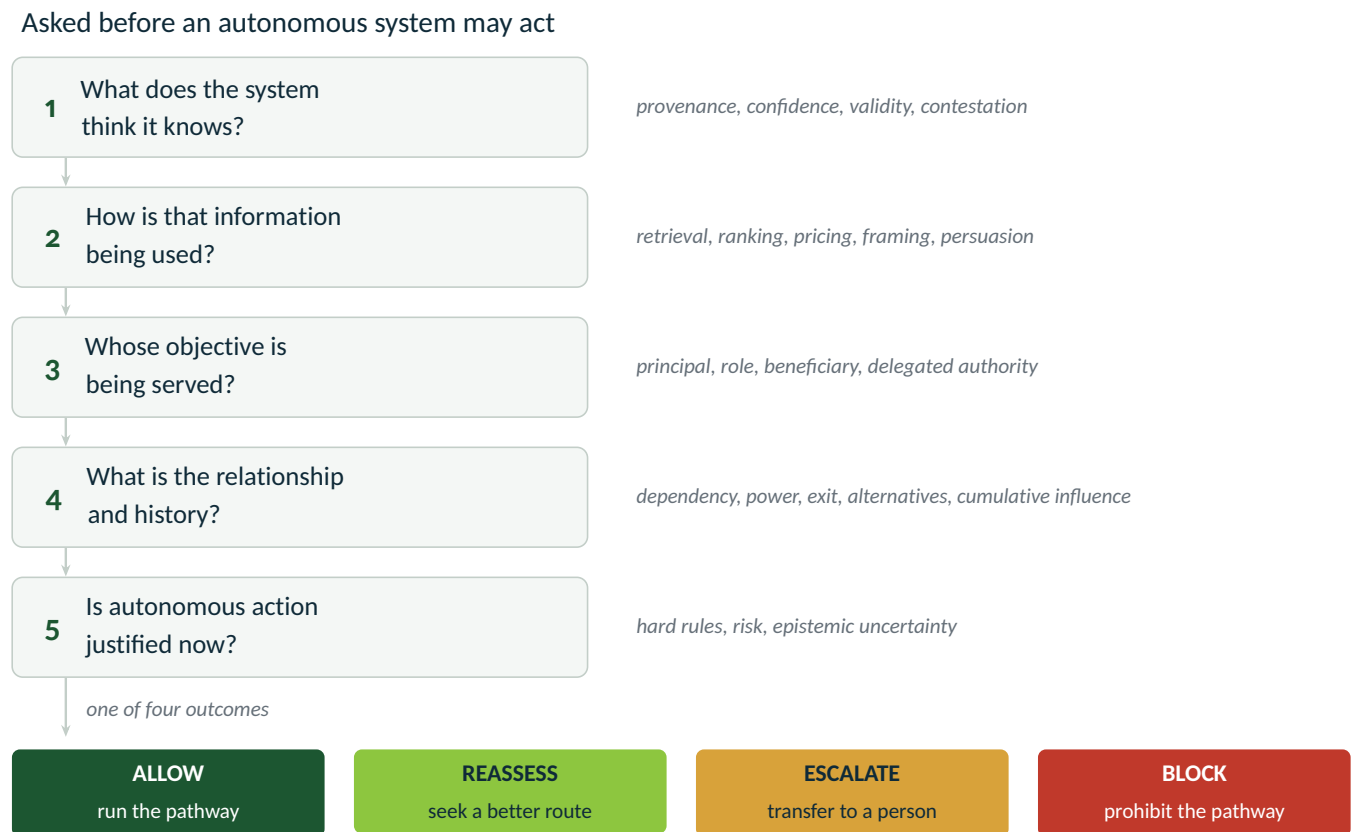
The Person-Model Use Gate (PMUG) asks a local question before the final action is authorized: may this person-model

information be used in this way for the claimed purpose and beneficiary? The important point is not the acronym or notation. It is the separation between possession of information and permission for that information to influence a decision.

This is also where contestability stops being abstract. A person can dispute that an inference is factually wrong. But they can also say something different: ‘even if this inference is accurate, you are not authorized to use it to manipulate price or pressure me.’ Contestable-AI research already treats responsiveness to dispute as a design requirement (Alfrink et al., 2023). The DSL proposes a further step: a valid challenge can suspend a purpose, require re-verification, shorten retention or force future use to human review. The DSL does not assume that self-report always outweighs other evidence. It requires the conflict, and the authority for resolving it, to be made explicit.

4 How the Dignity Safety Layer works

The DSL is a staged runtime architecture. For a general readership it can be understood through five questions asked before an autonomous system is allowed to proceed (Figure 3, Table 2).



REASSESS returns the pathway to question 1 with less intrusive evidence.

Figure 3. The five questions and the four outcomes. Each question is asked of a specific consequential decision before the system is allowed to act autonomously; the pathway is the pair of an information use and the action it helps select.

Question	What the system checks	Why it matters
1. What does the system think it knows?	The relevant person-model feature: observed, declared or inferred; its provenance, confidence, validity and contestation status.	A prediction is treated as a fallible representation, not an unquestionable fact.
2. How is that information being used?	The consequential operation: retrieval, filtering, ranking, pricing, framing, persuasion or another auditable decision step.	The same information can be legitimate for one use and exploitative for another.
3. Whose objective is being served?	The principal, represented role, beneficiary, operative objective, delegated authority and material conflicts.	A “personal assistant” may actually be optimizing a merchant, employer or institutional objective.
4. What is the relationship and interaction history?	Dependency, power, alternatives, exit, reversibility and cumulative influence across time.	An intervention acceptable in an ordinary choice setting may be coercive in employment, credit or public benefits.
5. Is autonomous action justified now?	Applicable hard rules, risk, epistemic uncertainty and action-level constraints.	The system can proceed, re-plan, transfer authority or refuse the pathway.

Table 2. The five questions the DSL asks before an autonomous system may act.

Four possible outcomes

ALLOW means the pathway, this use of information for this action, is authorized to run autonomously. REASSESS means the system should try a less intrusive or better-supported route: ask the person directly, remove a sensitive feature or use a declared preference instead. ESCALATE means the uncertainty or the stakes are too high for the machine to decide alone; the case must reach a legitimate human decision-maker. BLOCK means a scoped rule prohibits the pathway outright.

WHY NOT JUST USE A PENALTY SCORE?

A finite penalty can always be outweighed by enough task utility. For requirements that are deliberately designated non-tradeable, the DSL uses an admissibility boundary: utility is optimized only inside the set of authorized pathways.

The architecture also treats uncertainty as distinct from risk. A system may estimate a low probability of a violation while still being unsure whether its detector applies in the present context. Missing evidence, out-of-distribution cases or unstable explanations can therefore justify REASSESS or ESCALATE even when estimated risk is low. This is intended to prevent confidence in an incomplete model from silently expanding machine authority.

Where the relevant use of information cannot be exposed or mediated, the DSL can promise less, and must say so. This is an important limitation for end-to-end neural agents. High-consequence implementations may need purpose-bound memory services, controlled tools, information-flow labels or other architectural separation rather than relying on post-hoc explanations.

5 Why relationship and history matter

Human agency is not exercised in a vacuum. Dependency, power, constrained alternatives and lack of exit can change the practical meaning of an intervention. Relational-autonomy scholarship makes this explicit, while republican accounts of non-domination focus on exposure to arbitrary power (Mackenzie & Stoljar, 2000; Pettit, 1997). The DSL therefore represents the relationship itself rather than assuming that the same content means the same thing everywhere.

A persuasive recommendation between optional entertainment products may be relatively low consequence. The same optimization strategy in debt collection, employment termination, insurance, healthcare or access to public benefits can be much more serious because the system or institution controls something the person depends upon. The DSL does not turn this into a universal ‘power score’. Instead, domain rules specify which relational variables matter and what they change.

Two comparisons make this concrete. A recruitment agent infers that a candidate is anxious about a gap in their employment history and uses that inference to shape the timing or framing of a lower salary offer in order to increase the probability of acceptance. The inference may be accurate, and the action, sending an offer, is unremarkable. What changes the assessment is that the employer controls access to the role, the candidate has no comparable alternative in view, and the use serves the employer’s objective rather than the candidate’s.

A public-benefits assistant infers that a claimant is confused by an eligibility rule. Used to simplify the explanation and offer a callback, that inference supports the person’s agency. Used to steer the claimant toward withdrawing or not pursuing the claim, with the operative objective of reducing caseload, the same inference exploits the dependency that the supportive use respected. The action is a message in both cases. The pathway and the objective are not the same.

Time introduces another difficulty. Harmful influence can be cumulative even when each individual action appears benign. One reminder may be acceptable. Repeated urgency cues, selective omission of alternatives and progressively personalized pressure can alter the choice environment over several interactions. This is why the DSL includes trajectory-level assurance in addition to local action checks.



Figure 4. An illustrative trajectory mechanism. Each individual step remains below a local threshold, while cumulative influence can cross a governed review boundary. The values are illustrative, not measures of dignity or manipulation.

Trajectory monitoring creates its own privacy problem because detecting cumulative influence requires memory. The architecture therefore favors minimized, purpose-limited state, such as counters or signed events, over indefinite retention of raw conversations where possible. A credible design must report this privacy cost rather than treating more surveillance as an automatic route to more safety.

6 What the synthetic experiments show, and what they do not

Three synthetic experiments were carried out. Their purpose is not to measure dignity. They test whether the proposed controls behave the way the architecture predicts. The distinction matters because a synthetic benchmark can demonstrate logic. It cannot establish real-world social validity.

How the experiments were run. All three are simulations, not measurements of people. Experiment A draws 30,000 synthetic episodes with a vulnerability parameter from a Beta(2.2, 2.2) distribution, a high-stakes indicator on 72 per cent of episodes and six candidate actions; the conventional action constraint excludes actions with estimated risk at or above 0.42, and the DSL gate additionally excludes actions whose epistemic uncertainty reaches 0.68. Experiment B holds the external action fixed across 30,000 cases while varying which of four person-model uses selected it, two of the four being prohibited by the synthetic policy. Experiment C generates 8,000 ten-step sequences whose local influence increments stay below 0.29, with a decayed accumulator (decay 0.90, threshold 0.72) defining the trajectory label. Every experiment uses the same fixed random seed (20260814). The violation rules, the distributions and the thresholds are deliberately simplified assumptions chosen to exercise the architecture, not calibrated estimates of anything in the world.

Experiment	Question tested	Main finding	How to interpret it
A. Action gating	What happens when explicit constraints and uncertainty limit ordinary utility maximization?	The DSL-style gate did not materially reduce toy violations beyond the conventional action constraint in this simulation (0.403% versus 0.383% on 30,000 episodes). What it changed is how often the system withheld autonomous action: about 0.5% of cases under the conventional constraint against about 1.8% under the DSL gate, at marginally lower synthetic utility.	The measurable consequence of treating uncertainty as a reason to withdraw authority is more escalation and abstinence, not a demonstrated reduction in violations. A design trade-off, not evidence that the DSL is generally superior.
B. Same action, different use	Can an action-only monitor distinguish identical outputs selected through different person-model uses?	No. In the constructed benchmark the action-only gate has zero true-positive rate for prohibited use, while the PMUG can distinguish the pathways.	The contribution is representational: the benchmark holds action fixed while changing the information-use path.
C. Cumulative influence	Can a sequence cross a temporal condition even when every local event stays below a local threshold?	Yes, by construction; intervention varies materially with decay, window and threshold choices.	Shows why temporal settings are normative-operational assumptions that must be disclosed and tested.

Table 3. What the three synthetic experiments test, and how far each result reaches.

The ablation study adds action, person-model use, telos, relational state and trajectory components one at a time. Coverage rises, but only because the benchmark contains failure classes designed for those components, and the cost of intervention rises with it. This is useful as a test harness, but it does not tell us how common the relevant failures are in actual deployments.

RESEARCH INTEGRITY POINT

The strongest claim supported so far is architectural: a system that governs only external actions cannot represent every distinction that becomes visible when the informational pathway, purpose and relationship are also modeled. Whether this extra structure improves real safety enough to justify its cost remains an empirical question.

The next phase should therefore use expert- and participant-labeled cases, not detector-defined ground truth. Particularly valuable experiments would hold the external action constant while varying audited informational pathways and purposes; compare trajectory judgments for isolated actions versus full sequences; and measure false positives, false negatives, latency, privacy burden and human-review workload.

7 How the DSL relates to existing approaches

The DSL does not replace mature fields. It sits at their intersection. Constrained optimization and shielding already show how an optimizer can be kept inside a permitted region. Purpose-based access control and UCON already separate access from permitted purpose and ongoing use. Contestable AI already treats responsiveness to dispute as a design requirement. Meaningful human control already emphasizes the importance of legitimate human authority. The DSL's contribution is therefore deliberately narrow.

Existing approach	What it already does well	What the DSL adds or emphasizes
Safe RL / shielding / runtime assurance	Constrains actions or trajectories against specified safety rules.	A dignity-oriented semantic layer that includes person-model use, purpose and relational context.
Purpose-based access control / UCON	Controls access and continuing use according to purpose, obligation and condition.	Treats an internal, consequential use of a person-model as a runtime authorization event linked to the downstream action.
Privacy / contextual integrity	Governs appropriate information flows in social context (Nissenbaum, 2010).	Highlights that disclosure can be permitted while exploitation of the same fact for persuasion or pricing may remain prohibited.
Contestable AI / recourse	Provides mechanisms for challenge and response.	Proposes that valid contestation can change future feature-use permissions, not only the outcome of one appeal.
RLHF / Constitutional AI	Shapes model behavior through feedback or principles.	Adds an external authorization boundary for selected high-consequence operations; it is complementary rather than a substitute.

Table 4. How the DSL relates to established approaches.

Purpose-based access control and UCON are the sharpest challenge to the originality of the DSL. The contribution survives only if the DSL governs something more specific than ordinary data access: the point where person-model information actually shapes an agent's decision, linked to the action that follows and to the relationship and history around it. Where existing usage-control or information-flow systems already express the needed condition, the DSL should reuse them rather than duplicate them (Byun & Li, 2008; Park & Sandhu, 2004).

Dignity-specific technical modeling also has precedent. Deniz and Stiefenhofer (2026), for example, incorporate a dignity-linked constraint in an AI-driven hiring model. The proposed contribution here is therefore not 'the first

dignity constraint in AI'. It is bringing those threads together at runtime (how the person-model is used, whose purpose it serves, the relationship and history, how uncertain the system is, and whether the person can challenge it) around machine action taken on delegated authority.

8 Risks, limitations and the research program

A dignity-oriented control layer can itself create risks. The most serious is normative overreach: an institution may encode its own interpretation of dignity as if it were universal. The DSL responds by requiring scoped rules with visible provenance and approving authority, but this does not solve the deeper legitimacy question. A technically correct system can still enforce an unjust policy.

Paternalism is a related danger. If an AI labels someone 'vulnerable', the label can itself become a reason to restrict that person's choices. The architecture therefore treats vulnerability inferences as provisional and purpose-bound. Wherever possible, relational context should first constrain the power of the system, for example by preventing exploitation, before it becomes a reason to override the individual.

Other limitations are practical. Person-models can be wrong or stale. End-to-end models may use distributed representations that cannot be mapped cleanly to an auditable use. Agents may reconstruct prohibited attributes from proxies or learn to operate just below thresholds. Trajectory monitoring increases privacy burden. Escalation can overwhelm human reviewers. Institutions can misstate their objectives or configure permissive rules. Formal verification can prove that encoded policy is enforced, but it cannot prove that the policy itself is morally legitimate.

What needs to be tested next

This brief treats the DSL as a falsifiable research program rather than a completed paradigm. Six questions deserve direct testing: whether action-only controls miss information-use distinctions; whether explicit telos provenance exposes conflicts missed by action monitoring; whether relational variables systematically change judgments of permissibility; whether trajectory monitors identify cumulative patterns missed locally; whether contestation changes subsequent model use; and whether non-tradeable prohibitions can be enforced without converting dignity into a scalar reward.

WHAT WOULD COUNT AS SUCCESS?

Not proving that a machine has "understood dignity." Success would mean showing that a clearly specified person-model governance layer identifies and prevents a meaningful class of unauthorized uses that simpler mechanisms miss, at acceptable cost, and that it can be governed, challenged and audited.

9 Conclusion

Personalized autonomous AI changes an old governance question. We are no longer deciding only what information organizations may collect or what actions machines may take. We must also decide how what a system knows about a particular person is allowed to shape what it does. The Dignity Safety Layer offers one way to make that authority visible and bounded. Its core proposition is modest but consequential: the fact that a system can model a person does not, by itself, give it unrestricted authority to act through that model. Whether the proposal earns its place will depend on two things: how it performs against simpler controls in real empirical comparison, and whether legitimate institutions can specify, oversee and contest the rules it enforces.

Appendix. A five-question check for a design review or procurement

The five questions in Section 4 can be used directly as a review instrument. Ask each one of a specific consequential decision the system makes, and record the answer that the system's own documentation and logs can support rather than the answer the design intended.

- 1. What does the system think it knows about this person, and where did each element come from?** Record which features are declared, observed and inferred, and whether provenance, confidence, validity period and contestation status are recorded at all. A feature with no provenance cannot be governed.
- 2. How is that information used in this decision?** Name the operation: retrieval, filtering, ranking, pricing, framing, persuasion, termination. If the pathway cannot be named, the use cannot be authorized, and the assurance claim has to be narrowed to what can be observed.
- 3. Whose objective does the use serve?** Record the principal, the role represented to the person, the intended beneficiary, the operative objective and any material incentive. Note every case where the represented role and the operative objective differ, and what is disclosed.
- 4. What is the relationship, and what has happened before?** Record dependency, control over a resource the person needs, realistic exit, alternatives and reversibility, and whether cumulative influence across repeated interactions is monitored at all.
- 5. Is autonomous action justified now?** Record which rules apply, who approved them, what the system does under uncertainty, and which of ALLOW, REASSESS, ESCALATE or BLOCK it would return. If the answer is always ALLOW, the layer is not doing anything.

Declarations

Ethics: This brief reports no work with human participants and processes no personal data. All computational results come from synthetic simulations: the episodes, sequences and person-model features are generated, and no individual's data was collected, used or inferred from.

Data and code availability: The parameters, thresholds and random seed behind the simulations are stated in Section 6. The simulation code and the benchmark-generation materials are not published.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The work was undertaken by the author with institutional support from Cognovation.

Competing interests: The author is affiliated with Cognovation and is the inventor/applicant on a patent application filed on 13 August 2026 relating to the broader EthOS research and technology programme, of which the Dignity Safety Layer forms part. The author may therefore have a future commercial interest in technologies related to the subject matter of this brief. This work reports a proposed research framework and synthetic experiments and does not evaluate a commercial product. No other competing interests are declared.

About this brief: This brief sets out the Dignity Safety Layer for the people who would have to authorize, specify and contest its rules: policy and governance staff, compliance and ethics functions, and those procuring or reviewing autonomous systems. The framework places normative interpretation upstream of the runtime gate, so it depends on institutions being able to state narrow, reviewable rules. This document is written to make that possible without the formal apparatus, and it sets out the architecture as five questions and four decision outcomes that can be applied to a design review or a procurement.

Status of the evidence: The computational results reported in Section 6 are synthetic mechanism-validation results. They test whether the proposed controls behave as designed. They are not measurements of human dignity, and no real-world, expert-labelled or participant-labelled validation has yet been carried out. The parameters and the fixed random seed behind each experiment are stated in Section 6; a fuller technical manuscript is held by the author and may be made available for scholarly review where appropriate. The simulation code and benchmark-generation materials are not published.

References cited

- Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable AI by design: Towards a framework. *Minds and Machines*, 33, 613–639. <https://doi.org/10.1007/s11023-022-09611-z>
- Byun, J.-W., & Li, N. (2008). Purpose based access control for privacy protection in relational database systems. *The VLDB Journal*, 17, 603–619. <https://doi.org/10.1007/s00778-006-0023-0>
- Council of Europe. (2024). *Framework convention on artificial intelligence and human rights, democracy and the rule of law* (CETS No. 225). Council of Europe.
- Deniz, C., & Stiefenhofer, P. (2026). The algorithmic conscience: Reconciling AI efficiency with human dignity. *Journal of Chinese Economic and Business Studies*. <https://doi.org/10.1080/14765284.2026.2640533>
- European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act)*. European Union.
- Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)
- Kaptein, M. C., Markopoulos, P., de Ruyter, B., & Aarts, E. H. L. (2015). Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77, 38–51. <https://doi.org/10.1016/j.ijhcs.2015.01.004>
- Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.
- McCrudden, C. (2008). Human dignity and judicial interpretation of human rights. *European Journal of International Law*, 19(4), 655–724. <https://doi.org/10.1093/ejil/chn043>
- Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
- Park, J., & Sandhu, R. (2004). The UCONABC usage control model. *ACM Transactions on Information and System Security*, 7(1), 128–174. <https://doi.org/10.1145/984334.984339>
- Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford University Press.
- Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 15. <https://doi.org/10.3389/frobt.2018.00015>
- Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2). <https://doi.org/10.14763/2019.2.1410>
- Teo, S. A. (2023). Human dignity and AI: Mapping the contours and utility of human dignity in addressing challenges presented by AI. *Law, Innovation and Technology*, 15(1), 241–279. <https://doi.org/10.1080/17579961.2023.2184132>
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence* (Adopted 23 November 2021). UNESCO.
- United Nations. (1948). *Universal declaration of human rights*. United Nations.
- Waldron, J. (2012). *Dignity, rank, and rights*. Oxford University Press.

Further reading

Background sources that inform the framework but are not cited in this brief.

- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30, 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, 35, 957–967. <https://doi.org/10.1007/s00146-020-00950-y>
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Raji, I. D., Smart, A., White, R. N., Zaldivar, M., Barnes, P., Hutchinson, B., Smith-Loud, J., Theron, D., & Mitchell, M. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*.

Copyright and publication

© 2026 Mazhar Ali Bari. Published by the International Center for Artificial Intelligence Research and Ethics (ICAIRE), Riyadh, Saudi Arabia. All rights reserved.

Review

This paper was reviewed by the ICAIRE Scientific Committee and approved for publication by the Committee Chair, subject to revisions, which were incorporated by ICAIRE in accordance with the ICAIRE Research Policy.

Disclaimer

The views expressed are those of the author and do not necessarily reflect the positions of ICAIRE or UNESCO.

Suggested citation

Bari, M. A. (2026). A dignity safety layer for autonomous AI (ICAIRE Study Brief). International Center for Artificial Intelligence Research and Ethics. <https://doi.org/10.68480/ANNN8336>