# Robustness by Design: Interface Contracts for AI Control in High-Stakes Physical Systems

**Vacslav Glukhov**
vg@nextfusion.org
Next Step Fusion S.a.r.l., Luxembourg

**Georgy Subbotin**
gs@nextfusion.org
Next Step Fusion S.a.r.l., Luxembourg

**Maxim Nurgaliev**
mn@nextfusion.org
Next Step Fusion S.a.r.l., Luxembourg

## Abstract

Fusion power plants and other high-stakes physical systems demand levels of robustness and safety unattainable by purely data-driven or purely physics-based control. Hybrid architectures that integrate both paradigms are therefore not a design choice but an operational necessity. We argue that their reliability depends on explicit *interface contracts* — formal specifications of how heterogeneous modules exchange information, share responsibility, and recover from uncertainty or model failure. Such contracts define domains of applicability, fallback and handover rules, and accountability mechanisms, transforming hybrid control into an auditable and certifiable framework. Using confinement-state identification in fusion plasmas as an illustration, we argue that robustness in AI-driven control arises from disciplined interface design rather than algorithmic novelty, enabling safe deployment of learning systems that act within the material world.

## 1 Introduction: plasma control challenge in fusion power plants

The recent surge of private investment in fusion research is fueled by the demand for AI infrastructure and fusion's promise of sustainable, abundant, and clean power. Controlling a fusion device requires genuinely complex control systems, not merely complicated ones. Data-driven methods open new opportunities in this domain, but they also introduce new challenges — especially in safety-critical, high-stakes settings.

Fusion plasmas exhibit highly nonlinear dynamics, and their physics remains an active area of experimental and theoretical research. In research environments, diagnostics and control systems are built for flexibility and versatility. They utilise diverse and often non-orthogonal sensors and actuators, often rely on human operators, and primarily focus on short-pulse scenarios. Plasma disruptions in this context are typically manageable events, offering opportunities to study plasma stability.

By contrast, fusion power plants must prioritise efficient and reliable power production. Their diagnostics and control systems are designed for robustness and maintainability, with sensor choice severely constrained by vulnerability to heat and radiation. Automation is essential for sustained steady-state operation, and plasma disruptions can result in severe inefficiencies — or even catastrophic failures. Moreover, the design and operation of fusion power plants are subject to multiple, often conflicting objectives (see e.g. (Lennholm et al., 2024a), (Lennholm et al., 2024b)).

Deep reinforcement learning (DRL) with scalar rewards and monolithic controllers has shown promise in research settings (Degrave et al., 2022; Subbotin et al., 2025). In fusion power plants, however, the complexity of plasma dynamics, diagnostic constraints, and stringent safety and performance

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

requirements expose the fundamental limitations of monolithic, purely data-driven control. Models trained on historical data can only generalise within regions of dense data support; beyond these, their behaviour becomes unpredictable. Under sensor degradation, noise, or unobserved regimes, they offer no guarantees of physical validity or safe operation.

Neither physics-based nor data-driven paradigms alone can provide the necessary coverage, adaptability, and reliability. Physics-based control cannot fully capture unmodelled effects or diagnostic degradation; data-driven control cannot operate safely in low-support regions of the state space where data are sparse and models are extrapolating. This complementarity renders heterogeneous architectures not a design choice, but a structural necessity for high-dimensional, nonlinear systems subject to uncertainty and regime drift.

To function safely, such architectures must include explicit *interface definitions* that specify how modules interact and under what conditions their outputs can be trusted, deferred, or overridden. Robustness does not arise from coexistence but from precisely defined operational boundaries and coordination logic. In plasma control, these interface contracts delineate domains of applicability, define fallback and handover rules, and establish traceable accountability, ensuring that data-driven components strengthen rather than compromise the reliability of physics-based control.

## 2 Data-driven vs physics-based models

Two implicit assumptions underlie much of modern data-driven modelling (Cotton, 2022). The first is that data are effectively infinite — that past observations densely cover the relevant state space. The second is an assumption of regularity: temporal regularity (the past predicts the future) and state-space regularity (behaviour in the vicinity of a state approximates behaviour at the state itself). Sharp boundaries, discontinuities, and regime changes common in real physical systems often violate these assumptions.

Within regions of dense data support, data-driven models can efficiently represent complex nonlinear mappings between inputs and outputs. However, their reliability degrades rapidly in low-support regions, where training data are sparse or absent, and where extrapolation rather than interpolation dominates. Such models can reproduce observed dynamics but lack intrinsic guarantees of physical validity, conservation, or stability when operating outside their training distribution. Their sensitivity to data loss or contamination and their limited decision traceability make them difficult to deploy directly in high-stakes, safety-critical control systems.

Physics-based models, in contrast, are explicitly grounded in known laws and conservation principles. They encode couplings and constraints among state variables through governing equations and inequalities, ensuring internal consistency within their domain of applicability. They naturally separate spatial and temporal scales and express quantities in dimensional form, which makes assumptions and limitations explicit. When validated, these models offer interpretability, stability, and predictability. However, their fidelity depends on simplifying assumptions, closure relations, and empirically calibrated coefficients that restrict their range of validity. As system complexity increases, capturing all nonlinearities, multiscale couplings, and material effects becomes infeasible.

Efforts in physics-informed machine learning (PIML/SciML) and surrogate modelling aim to bridge this gap. They typically seek computational efficiency — replacing or accelerating physics-based solvers through data-driven approximations that remain consistent with underlying theory (Quarteroni et al., 2025; Lam et al., 2023; Li et al., 2024; Galletti et al., 2025; Drgona et al., 2025). These approaches have shown promise in contexts where physics is well understood and data are abundant. In fusion-relevant control, however, diagnostic limitations, data sparsity, and evolving plasma regimes restrict their direct applicability.

Consequently, neither paradigm alone can meet the requirements of sustained and safe operation in fusion pilot plants. Physics-based models provide structure and internal consistency but limited adaptability; data-driven models provide adaptability but limited reliability in low-support or unobserved regions of the state space. Hybridisation is therefore not optional but necessary — the only viable route to combine physical consistency with data-driven responsiveness.

Yet hybridisation alone is not sufficient. Without well-defined interaction rules, heterogeneity can introduce new failure modes, obscure accountability, and erode robustness. We propose the notion of *interface contracts* as a unifying principle: explicit agreements that define when and how modules may operate, handover authority, or defer to one another.

The central challenge thus shifts from model development to *interface design* — specifying how heterogeneous modules exchange information, allocate responsibility, and maintain operational integrity under uncertainty and regime drift.

# 3 Interface contracts

Hybrid architectures that combine physics-based and data-driven components require more than coexistence: their interactions must be explicitly defined. Interface contracts provide a formal mechanism for doing so. They specify how modules exchange information, under what conditions their outputs remain valid, and how control authority is transferred when those conditions change. In complex control systems such as fusion power plants, such formalisation is not a refinement but a prerequisite for operational safety, reliability, and verification.

Interface contracts translate qualitative notions of validation and authority into enforceable operational logic. They do not replace physical or statistical modelling but regulate their interaction — defining when a data-driven estimator may augment or override a physics-based model, and when it must defer. By codifying these relationships, contracts ensure predictable hybrid interaction: they make hybrid control systems explainable, auditable, and resistant to silent failure.

## 3.1 Essential elements

The essential elements of interface contracts address three interdependent aspects of hybrid control: domains of applicability, fallback and handover, and accountability.

- **Domains of applicability.** For physics-based models, the domain of applicability is often intrinsic — for example, the validity of a stationary approximation for a given plasma configuration. Data-driven models, in contrast, rarely specify such limits: the transition between regions of dense data support and sparse or unobserved areas of the state space is continuous and ill-defined. A contractual specification of the operational domain must therefore be imposed externally, through admissible input ranges, uncertainty thresholds, or verified performance envelopes.
- **Fallback and handover.** Once applicability domains are defined, contracts must also specify how control authority is transferred when those boundaries are crossed. Uncontrolled extrapolation or silent model failure is unacceptable in reactor-scale control. Fallback logic must include explicit conditions: if inputs leave the admissible domain or uncertainties exceed thresholds, responsibility is transferred to another module, typically from data-driven to physics-based, or ultimately to heuristic safeguards. This ensures controlled and traceable degradation rather than uncontrolled failure.
- **Accountability.** Contracts ensure that each control action is traceable to the module responsible under clearly defined conditions. Such attribution enables post-event reconstruction, supports verification and auditing, and clarifies the operational context under which each decision was made.

## 3.2 Supporting elements

In addition to the essential components, certain supporting elements can reinforce the reliability and auditability of interface contracts. Among these, uncertainty quantification and explainability hooks provide complementary mechanisms for monitoring confidence and maintaining traceability within hybrid control systems. They do not define control logic directly but supply the diagnostic information required to enforce and verify contractual boundaries.

- **Uncertainty quantification.** Contracts may require models to emit estimates of confidence, variance, or error bounds alongside their primary outputs. These signals characterise the reliability of predictions and can be used to trigger contractual responses — for example: "fall back to the physics-based model if the confidence score drops below threshold." Quantified uncertainty strengthens domain enforcement, providing a principled mechanism for fallback or handover and reducing the risk of uncontrolled extrapolation. It also enhances traceability by documenting not only what decision was made, but how certain the system was at the time.

- **Explainability hooks.** Contracts may also specify metadata that make outputs interpretable and traceable without exposing internal model details. Examples include provenance identifiers (indicating which module produced the output), validity flags, or out-of-domain indicators. Such hooks enable the supervisory layer to apply contractual checks — for instance: "accept the data-driven estimate only if the out-of-domain flag is false." This prevents the uncontrolled use of invalid outputs and provides an audit trail for post-event analysis and regulatory verification.

Together, these supporting mechanisms provide the observability required for contractual enforcement. They ensure that hybrid control systems remain both transparent and verifiable, even as individual models evolve or recalibrate over time.

The principles outlined above remain abstract until instantiated in a concrete control scenario. To illustrate how interface contracts can be implemented in practice, we consider the problem of confinement-state identification in tokamak plasmas. This example demonstrates how physics-based and data-driven models can complement each other within a well-defined contractual framework — how operational domains, fallback conditions, and accountability links can be formalised to ensure reliable and auditable control in fusion-relevant environments.

## 4 Illustrative example: interface contracts in hybrid control

The operation of interface contracts can be illustrated using a generic hybrid-control scenario that combines a physics-based model and a data-driven estimator within a supervisory control loop. The physics-based model ensures consistency with known physical constraints and provides reliable predictions under a wide range of conditions. The data-driven estimator delivers fast local corrections or forecasts where sufficient data are available. Interface contracts define how these two components cooperate: (a) they specify the operational domains of each model, (b) establish quantitative thresholds for fallback or handover when uncertainty grows or data become unreliable, and (c) ensure that all actions taken by the controller are traceable to the responsible module through metadata and version identifiers.

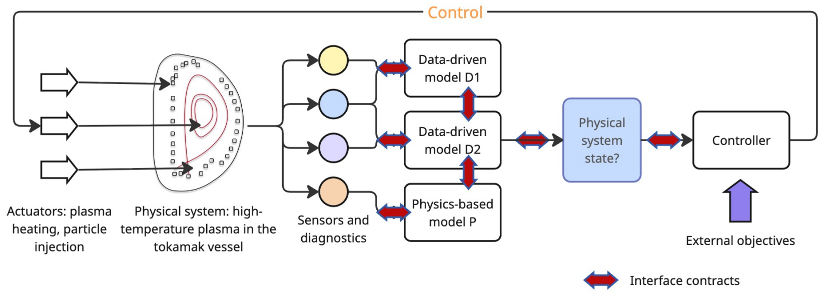

Figure 1: Hybrid-model-control schematic. Two data-driven estimators of the physical state and a physics-based model operate within a control loop. Interface contracts define operational domains, specify fallback or handover conditions between diagnostics and models, and between the models themselves, and provide accountability links to the controller system. Plasma state control logic is used as an illustration.

This structure can be instantiated across many physical-control domains. In fusion plasma control, for example, data-driven confinement-state classifiers analysing high-resolution features of temperature and pressure profiles can detect transitions between L- and H-modes with high precision when diagnostic and training coverage are sufficient (Clark et al., 2025). When diagnostic signals degrade or confidence drops below the contractual threshold, control authority is transferred to a physics-based model that infers confinement state from coarser but more reliable profile estimates. The physics-based model, while less precise in steady operation, remains valid across a broader operational space and preserves physical consistency under sparse data. Interface contracts mediate this interaction: the classifier operates only within its verified domain, while explicit fallback and accountability rules ensure that the overall control loop remains both adaptive and verifiable under uncertainty and regime drift.

## 5 Discussion: broader implications

The confinement-state example illustrates a broader truth: hybrid architectures are not a design preference but a necessity wherever physical complexity exceeds the reach of either data-driven or physics-based methods alone. In such systems, performance and safety depend not only on model quality but on the discipline with which boundaries of trust are defined and enforced. Interface contracts formalise this discipline, transforming ad hoc hybridisation into an auditable design framework.

This idea parallels the *design-by-contract* paradigm in software engineering, which improved predictability and traceability in large software systems by making preconditions, postconditions, and invariants explicit. Yet the analogy is incomplete. In physical control systems, contracts must operate in continuous, uncertain, and safety-critical environments. They cannot rely on binary assertions or static exceptions. Instead, they must handle graded confidence and physical consequences of failure. Where software contracts guarantee logical consistency, control contracts must guarantee operational robustness.

The necessity of such structures extends well beyond fusion. Aerospace systems, autonomous vehicles, defense, and energy grids face a similar tension: data-driven modules handle complexities and offer unprecedented adaptability but lack intrinsic guarantees of validity, while physics-based models provide guarantees but lack sufficient flexibility and adaptability. Interface contracts introduce a formal mechanism for managing interaction and trust among heterogeneous models. They create a path toward certifiable AI-driven control in domains where failure is not merely an error but a material hazard.

The central challenge is therefore not modelling physical phenomena, but deploying learning systems that act within them. When machine learning components become part of operational control loops in the material domain, their behaviour must be bounded, interpretable, and recoverable under uncertainty. Interface contracts provide the structure for achieving this: they define how trust is allocated, how authority transfers when confidence decays, and how every control decision remains traceable to its source. In this sense, robustness becomes a property of system architecture rather than of any individual model — a prerequisite for deploying AI in high-stakes physical environments.

## Acknowledgments and Disclosure of Funding

The authors acknowledge helpful discussions with Aleksandr Kachkin (Next Step Fusion), and Randall Clack and Dmitri Orlov (Center for Energy Research, University of California, San Diego).

## References

R. Clark, V. Glukhov, G. Subbotin, M. Nurgaliev, A. Kachkin, M. Austin, and D. M. Orlov. Plasma confinement state classification via fpp relevant microwave diagnostics. *Nuclear Fusion*, 2025. under review.

P. Cotton. *Microprediction: Building an Open AI Network*. The MIT Press, 11 2022. ISBN 9780262371346. doi: 10.7551/mitpress/13636.001.0001. URL `https://doi.org/10.7551/mitpress/13636.001.0001`.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, and M. Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602:414–419, 02 2022. doi: 10.1038/s41586-021-04301-9.

J. Drgona, T. X. Nghiem, T. Beckers, M. Fazlyab, E. Mallada, C. Jones, D. Vrabie, S. L. Brunton, and R. Findeisen. Safe physics-informed machine learning for dynamics and control, 2025. URL https://arxiv.org/abs/2504.12952.

G. Galletti, F. Paischer, P. Setinek, W. Hornsby, L. Zanisi, N. Carey, S. Pamela, and J. Brandstetter. 5d neural surrogates for nonlinear gyrokinetic simulations of plasma turbulence, 2025. URL https://arxiv.org/abs/2502.07469.

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

M. Lennholm, S. Aleiferis, S. Bakes, O. Bardsley, M. van Berkel, F. Casson, F. Chaudry, N. Conway, T. Hender, S. Henderson, et al. Plasma control for the step prototype power plant. *Nuclear Fusion*, 64(9):096036, 2024a.

M. Lennholm, S. Aleiferis, S. Bakes, O. P. Bardsley, M. van Berkel, F. J. Casson, F. Chaudry, N. J. Conway, T. C. Hender, S. S. Henderson, B. Kool, M. Lafferty, H. F. Meyer, J. Mitchell, A. Mitra, R. Osawa, R. Otin, A. Parrot, T. Thompson, G. Xia, and t. S. T. null. Controlling a new plasma regime. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2280):20230403, 2024b. doi: 10.1098/rsta.2023.0403. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2023.0403.

H. Li, L. Wang, Y.-L. Fu, Z. Wang, T. Wang, and J. Li. Surrogate model of turbulent transport in fusion plasmas using machine learning. *Nuclear Fusion*, 65, 11 2024. doi: 10.1088/1741-4326/ad8b5b.

A. Quarteroni, P. Gervasio, and F. Regazzoni. Combining physics-based and data-driven models: advancing the frontiers of research with scientific machine learning. *Mathematical Models and Methods in Applied Sciences*, 35(04):905–1071, Mar. 2025. ISSN 1793-6314. doi: 10.1142/s0218202525500125. URL http://dx.doi.org/10.1142/S0218202525500125.

G. F. Subbotin, D. I. Sorokin, M. R. Nurgaliev, A. A. Granovskiy, I. P. Kharitonov, E. V. Adishchev, E. N. Khairutdinov, R. Clark, H. Shen, W. Choi, J. Barr, and D. M. Orlov. Reconstruction-free magnetic control of diii-d plasma with deep reinforcement learning, 2025. URL https://arxiv.org/abs/2506.13267.