Science Needs AI Data Stocktakes

A proof-of-concept for fusion energy

By Conor Griffin, Don Wallace, and Theo Brown

For 40 years, amid green pastures outside Culham, a small village in Oxfordshire, scientists and engineers toiled at the Joint European Torus. They were attempting to harness nuclear fusion, a force powerful enough to light the sun.

To create fusion, scientists and engineers must heat the nuclei of very light atoms with such intensity that they fuse, instigating a self-sustaining reaction that releases vast amounts of energy. The scale of the challenge is hard to fathom—at its most extreme, the Joint European Torus, or JET, was the hottest point in the solar system, hitting over 150 million degrees Celsius.

JET concluded in 2023, generating a record amount of energy in its final experiments. The project is now part of fusion's history but remains pivotal to its future. The growing number of organisations developing fusion reactors are drawing on JET's discoveries. The UK Atomic Energy Authority is advancing a national fusion facility, MAST-U, on the Culham site. This will serve as a test-bed for STEP Fusion, the UK's project to put fusion electricity on the grid, set to begin operations in the early 2040s.

But JET didn't just bequeath novel discoveries. It left behind massive troves of data. That raises a tantalising prospect: Could scientists use this data to train AI models that accelerate the path to fusion power?

This is possible, but challenging. Most JET data is raw and unvalidated. Many important insights are buried in scientists' logbooks. The data that does exist is not available open source or, generally, for commercial use. Changing this may require agreement from all of JET's original partners across Europe. One expert we interviewed called JET data a 'stranded asset'.

Such data predicaments are not specific to JET or to fusion, but apply across all of science, even though science is precisely the domain where AI could yield its greatest benefit to society. New breakthroughs and startups are emerging quickly, from protein design to material design. Scientists are also keen users of fast-improving AI coding agents. But a lack of high-quality data will dampen progress. In most disciplines, large, high-quality datasets like the Protein Data Bank, which underpinned AlphaFold, are absent.

The scientific community needs to tackle this problem, and there are promising signs. Late last year, the UK government launched an AI for Science strategy, which includes a new collaboration with Renaissance Philanthropy to identify priority datasets. The US government's Genesis Mission aims to train AI models and agents on federal scientific data. Google.org has a dedicated AI for Science fund, which can fund datasets and tooling.

These examples suggest that, if the scientific community can identify the data that AI needs, a range of actors could help to fund and deliver it.

This demands what we're calling **AI data stocktakes**. The concept is simple: interview leading experts in a given scientific field to understand the main opportunities to apply AI; the data obstacles; and the interventions that could make the biggest difference. Admittedly, some blockages, such as a paucity of engineers, are structural and will take years to fully resolve. **AI data stocktakes** should identify such challenges, but focus on projects that governments, companies and philanthropies could fund and implement within 1-2 years.

There are promising early efforts to map AI data gaps. But, to our knowledge, there are no concise, accessible documents that explain the AI opportunities in genomics, weather forecasting, and food security and convert them into a list of fundable data projects for policymakers and funders to pursue.[1]

In this essay, we offer a proof-of-concept. We interviewed 25 leading experts to create an **AI data stocktake for fusion**. We focus on the UK, but our analysis and recommendations could be taken up by funders anywhere in the world. Moving forward, we hope to support AI data stocktakes for other scientific disciplines and research problems.

I. Why fusion? Why now?

Fusion generates huge amounts of energy by *fusing* the nuclei of very light atoms, typically isotopes of hydrogen, such as deuterium and tritium.[2] Unlike nuclear fission, which generates energy by *splitting* the nuclei of very heavy atoms, no working fusion power plants yet exist.

To achieve fusion, scientists need to create and control *plasma,* a super-hot state of matter, in which the atoms have been stripped of their electrons and extreme heat and pressure are used to force the remaining nuclei to collide and fuse.

In 1957, the British physicist John Lawson formulated the conditions necessary for a self-sustaining fusion reaction. To achieve this, scientists need to reach the so-called 'triple product' where they get the plasma to be hot enough, dense enough, and confined for long enough that it produces more energy than it consumes.

Scientists are pursuing two main approaches to doing this, with very different physics, data, and AI opportunities. *Magnetic confinemen*t fusion uses massive magnets to suspend the plasma inside a donut-shaped machine called a 'tokamak', or a twisted bagel-shaped machine called a

[1] Fusion has several characteristics that make an AI data stocktake exercise tractable, including a relatively small and centralised research community and early efforts to build on, like the open-source FAIR MAST initiative and the IMAS data standardisation effort. Fields like genomics, weather forecasting, and food security look quite different, and so careful thought is needed on how to best scope AI data stocktakes in these fields. Nevertheless, we think they would be useful.
[2] Isotopes are atoms that have the same number of protons in their nuclei, but different numbers of neutrons. For example, hydrogen, deuterium and tritium all have one proton, but, as their names imply, deuterium has two neutrons and tritium has three.

'stellarator'. Inertial confinement fusion (often) uses high-energy lasers to implode tiny pellets of hydrogen, triggering short bursts of energy.[3]

We focus our data stocktake on magnetic confinement fusion, as this is the approach that the UK's STEP project is based on. It's also the approach used by Tokamak Energy, the UK's leading fusion power startup and Google DeepMind's fusion team, But as we return to below, several experts argue that the UK should also focus more on inertial confinement fusion, not least due to the US National Ignition Facility's recent record-breaking results.[4]

Why fusion? A safe, limitless, emission-free source of power

As a power source, fusion would be extremely energy-dense, releasing ~ten million times more energy, per kilogram of fuel, than coal.[5] Unlike coal, it would be essentially emission-free, with a close-to-limitless fuel stock. As such, it could help to address climate change, energy scarcity and unlock valuable innovations whose energy costs may otherwise be prohibitive, such as desalination to address freshwater scarcity. Fusion also poses fewer risks than fission because you cannot use its input fuels, in isolation, to create a nuclear weapon. A fusion power plant would also not pose the kind of meltdown risk, or the high-level nuclear waste, that fission can.[6]

From a scientific perspective, achieving fusion would yield a better understanding of the plasma that makes up 99% of the visible universe. From a societal innovation standpoint, it would lead to a supply chain rich in technologies that could be applied much more widely, from the magnets used to control the reactions to the robots used to maintain and decommission the facilities. Some of these benefits are already visible. Companies are exploring if the superconducting tape inside fusion magnets could be used to power AI data centres more efficiently. Drug discovery companies are using simulation codes from fusion research to model how atoms and molecules interact. The UK startup Astral Systems hopes to use the neutrons generated in fusion experiments to produce isotopes to diagnose and treat cancer.

Why now? Research breakthroughs, technology shifts and new funding

Quips about fusion being "always 20 years away" was always somewhat misleading. When researchers plot data from ~70 years of fusion experiments, it shows a fairly steady progression towards the 'triple product' goals of temperature, density, and confinement. In most fields, such progress would have solved major problems of interest decades ago. But fusion is an *extremely* hard problem. And the primary product is only attainable at the end of the line.

The end of the line *is* getting closer. The research breakthroughs continue, with scientists in Germany's Max Planck Institute, China's EAST facility, and France's WEST facility recently

[3] There are also other approaches to fusion, but the majority use tokamaks, stellarators or inertial confinement.

[4] In 2022, the National Ignition Facility achieved 'ignition' when they created a self-sustaining plasma that generated more energy than that which the lasers delivered to the target. Since then, NIF teams have steadily increased the amount of energy generated.

[5] Measuring and comparing energy density across fuel types is complex. See here and here for more details.

[6] There are caveats to the claim that fusion power would be essentially limitless, emission-free, and have no safety risks. One of the input fuels, tritium, is not widely available and scientists will need to use nascent 'blankets' to breed it from lithium. Certain parts of fusion reactors will become radioactive over time, although they can likely be recycled after ~50 years. Thermonuclear weapons use fusion reactions. But the weapons first require *fission* reactions and materials like enriched uranium and plutonium.

demonstrating how to maintain plasmas for longer durations—one of the keys to successful power generation. The underlying technology landscape has also changed. In addition to AI, the discovery of high-temperature superconducting magnets that can produce much stronger magnetic fields makes it easier to build much smaller and potentially cheaper tokamaks, more quickly, such as Commonwealth Fusion Systems' SPARC facility outside Boston.

Given its resource-intensity, fusion's traditional reliance on government funding has led to inevitable shortfalls and delays. In the past five years, a wave of private investment into startups like Commonwealth Fusion Systems and Tokamak Energy has transformed the sector. The Fusion Industry Association now identifies more than 30 companies pursuing fusion power generation, with some raising billion dollar deals and targeting commercially-viable pilot plants by the mid-2030s. This has injected momentum into the field, but also significant hype. In response, we need a clear view on the primary bottlenecks that AI can address.

II. How to accelerate fusion with AI

Fusion scientists and engineers want to *predict*, *control* and *understand* how plasma behaves. The challenge is that plasmas are highly complex and much of their underlying physics—from fluid dynamics to electromagnetics—remains poorly understood.

How to make progress? **One approach is to run experiments** to create plasmas in a physical reactor like the UK's MAST-U, and use sensors to measure its properties under different conditions. In doing so, scientists can validate their theories, reveal unexpected phenomena and test the hardware needed for power-plant-class devices. For example, how does altering the gas pressure or the shape of the magnetic field affect the start of a plasma? Does keeping the plasma in a certain shape avoid the damaging bursts of energy that can occur at its edge?

Building fusion reactors is extremely expensive and so they are few in number, with most researchers running their experiments at ~10 leading facilities worldwide. When they can get access to these facilities, researchers must choose from a huge array of experimental parameters and grapple with the limited real-time data available. Given the complex engineering involved, experiments often deviate off course, damaging components in the device and negatively affecting the data generated.

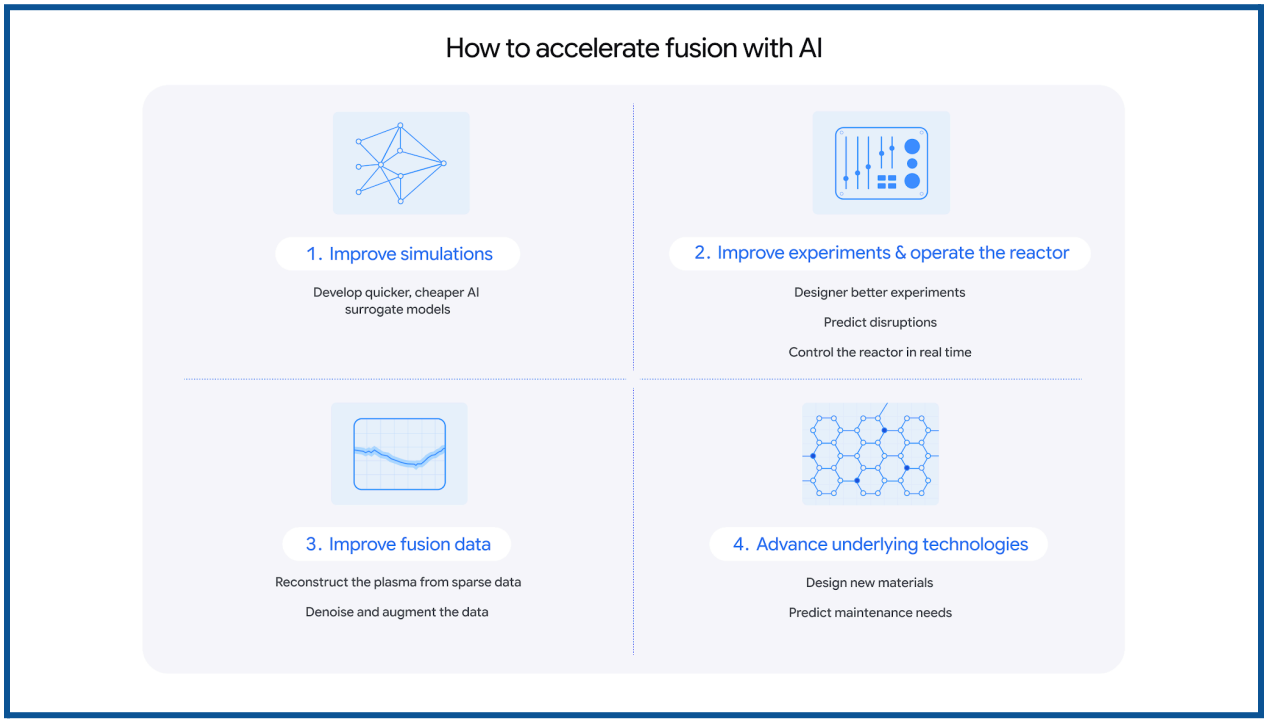
Given these challenges, fusion researchers also **run computer simulations** to help them better design and interpret their experiments, and to test variants of the many components needed for reactors. Researchers typically need to simulate a diverse range of phenomena, at very different scales, from the tiny, lightning-fast movements of electrons to the larger, slower evolution of the entire plasma. Using simulation codes to solve the equations governing these phenomena, with high fidelity, can be extremely computationally intensive. For simulations run on massive supercomputers, this may mean weeks. For those without such resources, simulations may take months. As a result, scientists make trade-offs, using assumptions and approximations to run their simulations more quickly and cheaply, but also less accurately.

The challenges don't stop there. When they try to bridge the worlds of simulations and experiments, researchers face the recurring *sim-to-real* challenge. They know that their theories, simulations and experiments are imperfect, with known and unknown limitations. But when a gap emerges between what a simulation suggests and what an experiment reports, it is often unclear where exactly the issue falls.

The AI opportunity

There is an opportunity to use AI in fusion, but also an *imperative*. This can sound like tech solutionism, but as one interviewee argued, it is more a testament to the awesome scientific and engineering complexity that fusion poses, which is likely impossible to solve without AI. The imperative is also economic. ITER is the world's flagship international fusion research collaboration, operating at a huge scale and offering major scientific benefits. Costs there can exceed US$1million per day. So delays matter.

The primary question is *where* and *how* to best use AI. In recent years, authors from MIT, the Clean Air Task Force, IAEA, FusionFest and the US Department of Energy have laid out some of the primary opportunities and early case studies.[7]



Opportunity 1: Improve simulations

Develop AI surrogate models

[7] Throughout this section, when we say 'AI' we generally mean deep learning (e.g. for AI surrogate models); large language models (e.g. for querying experimental logs); reinforcement learning (e.g. for controlling reactors); as well as more traditional statistical learning methods (e.g. Bayesian optimization or reconstructing plasma state from noisy measurements.).

To run fusion simulation codes more quickly and cheaply, researchers can use AI to develop *surrogate* models that emulate a code's predictions at a fraction of the computational cost. To create an AI surrogate, scientists run a simulation code many times. Each time, they vary the input parameters, such as the current in the magnetic coils. They then use the resulting dataset to train an AI model to predict the outputs of interest much more quickly.

For example, *plasma turbulence* describes the chaotic, swirling motion of particles that determines how well the plasma is confined—one of fusion scientists' 'triple product' goals. To simulate plasma turbulence, scientists typically use codes that make simplified assumptions so that they can be run more quickly. In a recent paper, researchers trained an AI surrogate for one such code, TGLF. This reduced the time needed to simulate an experiment for the UK STEP power plant from "several weeks" to "under 4 hours".

Moving forward, researchers could use AI to develop surrogates for more complex simulation codes like GENE and CGYRO, which make much fewer simplifications than codes like TGLF. This makes them more accurate, but often prohibitively expensive to run. Having AI surrogates for such codes would make simulations faster, but also more accurate.

In isolation, a better AI surrogate model for plasma turbulence will be of limited use. Rather, scientists need to stitch together many surrogates, each focussed on different aspects of a reactor, such as the 'scrape-off layer' where the exhaust heat hits a reactor's walls. This 'integrated model' would provide scientists with the capability to run a larger number of richer simulations, more cheaply, ultimately leading to much more powerful experiments and reactor designs. Google DeepMind's TORAX is an early example of open-source code, with an AI-native methodology, that enables multiple surrogates to be efficiently stitched together.

The field remains far from having a fully-functioning, integrated model that can simulate all aspects of a reactor with high fidelity. Rather, most AI surrogates are "one and done" efforts that result in a paper and some code, but quickly become out-of-date when scientists' understanding of the physics improves and they update their simulations.

To address this, the fusion community needs to shift to a system where simulation codes and AI surrogates are optimised, documented, and maintained—at scale. Among other things, this will require 'refactoring' simulation codes so that they can be run in parallel on modern chips, like GPUs. It will also mean automating the most labour-intensive aspects of running simulation codes, such as checking that it has converged on a stable solution or detecting errors early.

Opportunity 2: Improve experiments and operate the reactor

AI could help scientists to optimise the parameters in their experiments, predict disruptions that might knock the experiments off course, and ultimately help to automate the control of reactors.

Design better experiments

Most fusion experiments are a balancing act. Scientists must decide how to tune an array of parameters, from the electrical current in a reactor's coils to the input-valves that control the gas

levels. Researchers traditionally use simulations, human intuition and ongoing iteration to configure these parameters. AI could help them navigate the vast parameter space much more efficiently.

One approach is to use specialised techniques like Bayesian optimisation to predict the optimal parameters for the next experiments by learning from past experiments. By providing a probabilistic map of potential outcomes, researchers can strike a better balance between 'exploiting' settings that are known to work well and 'exploring' those that are more uncertain.

Researchers are also starting to use AI to design experiments in other ways. For example, researchers at the National Ignition facility used AI to predict how well their recent experiments would fare, overall. Scientists are also testing if LLMs can check and refine experimental protocols, such as how strong a laser needs to be to achieve a certain temperature.

Predict plasma disruptions

Disruptions describe the sudden catastrophic loss of plasma confinement. They end experiments, damage the reactors, and are one of the main obstacles to any future power plants. Disruptions can be triggered by a variety of factors, such as impurities in a reactor's wall and typically happen too fast for the control algorithms used in reactors to stop them.

Some computationally-expensive simulation codes, like M3D-C1 or JOREK, can model the full chain of events that lead to plasma disruptions, but they are not usually accurate enough to prevent disruptions in a reactor in real-time. As a result, researchers have created databases of past disruptions and tasked researchers with training AI models to predict them. These models have enjoyed success, predicting *historical* plasma disruptions with 99% accuracy. The next challenge is to develop models that can predict *future* disruptions on new, more powerful machines, and continually learn from the future data that these machines will provide.

If such AI models could be developed, scientists could use them to steer their experiments and help control future power plants. They could also help scientists to understand the underlying causes of disruptions, with researchers working on interpretability-driven approaches that automate the detection and identification of the events leading up to a disruption.

Control reactors in real-time

Fusion reactors operate in a real-time feedback loop, using an array of sensors to monitor the shape and temperature of the plasma and then adjusting the actuators, such as the magnetic coils that push and pull the plasma, accordingly. Traditionally scientists use classic control theory algorithms to determine the optimal 'policies' for how to do this. But these algorithms struggle to handle the chaotic, non-linear nature of millions of plasma variables interacting. Researchers are now training reinforcement learning agents to learn more effective control policies, including those that a physicist may never conceive of.

In 2022, researchers at Google DeepMind and the Swiss university EPFL demonstrated how an RL agent, based on a single neural network architecture, could control all 19 magnetic coils in a

tokamak. In 2024, another team of researchers showed how an RL agent could reduce the likelihood of tearing instabilities, a leading cause of plasma disruptions.

However, pure RL approaches can struggle to generalise beyond the scenarios encountered during training. To address this, scientists have developed 'hybrid' AI models that integrate some knowledge of physics. For example, researchers recently built a hybrid that learns how to safely shut down a plasma, while another team demonstrated a hybrid that combines aspects of traditional tokamak control with AI simulations of the more computationally-heavy parts.

Opportunity 3: Improve fusion data

As we expand on below, a lack of high-quality fusion data inhibits many opportunities to use AI. But AI could also help to improve fusion data in several ways.

Denoise and augment data

Fusion experiments are extreme environments. The intense heat and the chaotic nature of the plasma mean that the data that sensors pick up is often noisy or low quality. Some variables cannot be directly measured, and must be inferred, introducing additional sources of error. In response practitioners are training AI models to extract clean signals from this noisy data.

Scientists can also use AI to augment the data that is available. For example, the diagnostics in a fusion reactor also operate at different resolutions. Some may measure the plasma's temperature and density in ways that are highly accurate and granular, but slow. Others may do the inverse. Scientists at Princeton recently used AI to learn correlations between these diagnostics and create 'virtual diagnostics'. These virtual diagnostics can provide predictions for slower diagnostics but also ensure that data for those diagnostics continues, even if they get damaged—a capability that may be critical in future power plants.

Reconstruct the plasma from sparse data

The diagnostics and sensors that scientists use to generate data in a fusion experiment can't survive the most extreme parts of the reactor. This means that researchers can't directly measure all of the properties that they care about, such as the plasma's shape, pressure, and electrical current. So, traditionally, they feed the sparse data that they can gather into simulation codes like EFIT, that 'reconstruct' a more complete map of the data.

Researchers are developing AI surrogates to accelerate the most computationally-intensive aspects of these reconstructions. They are also using AI to help calibrate the simulation's outputs to the ground truth provided by the experimental data. Again, this involves creating synthetic diagnostics where the simulation provides predictions for what the sparse experimental data will be. If the predictions prove correct, it is a sign that the simulation is accurate. This process leads to both a more accurate reconstruction of the plasma, and a more accurate simulator.

Detect important events

Scientists often care less about the raw data from their fusion experiment or simulation, and more about certain events, such as when a disruption to the plasma began. Being able to accurately detect these events is critical to training AI models to predict them. Today, scientists often need to manually inspect graphs and plots—a difficult, laborious process. Scientists can use AI to automate parts of this process and to detect events that experimentalists may have missed, such as when diagnostics start to deviate from their expected behaviour.

Opportunity 4: Improve the underlying technologies
Achieving fusion will require a supply chain rich in technologies that could be applied more broadly. AI could help to accelerate them.

Design new materials
Designing the materials that fusion will require, such as chamber walls that can withstand extreme temperatures, is a grand challenge in and of itself, that could benefit many fields, from aerospace to energy storage. Today, the discovery of new materials is typically a slow and expensive trial-and-error process, where practitioners must identify candidate materials from a vast chemical space, predict their properties and test the top candidates in a lab.

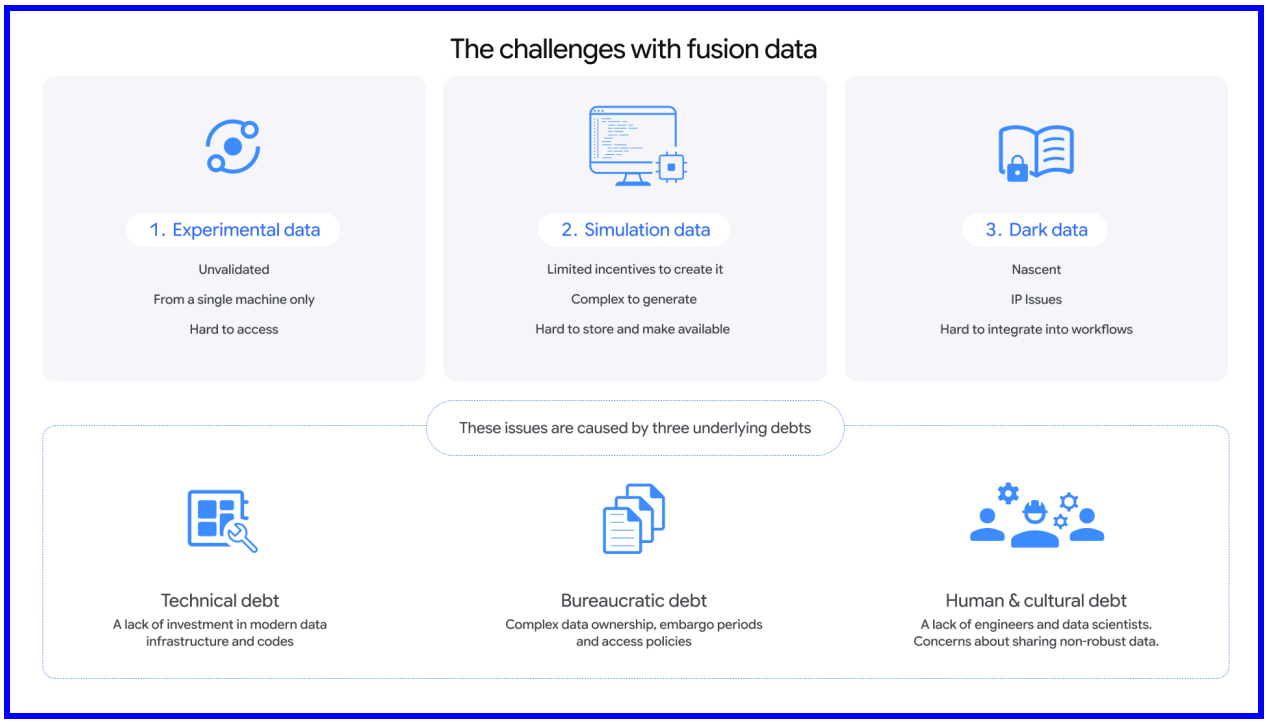
To accelerate this process, scientists could train AI surrogate models on data from slow, quantum-mechanics calculations to predict the forces between atoms millions of times faster. These surrogates could make it computationally tractable to run the kinds of large molecular dynamics simulations that are needed to assess a candidate material's real-world properties, like how strong or resistant to radiation it might be, over a fusion power plant's entire lifetime.

Predict maintenance needs
A typical fusion reactor may spend most of its time out of operation, due to different systems being unavailable or damaged. For example, JET faced regular issues with its power supplies and the beams that scientists used to heat the plasma—a major financial cost. AI systems could take in historical data from different diagnostics and learn the subtle signatures that indicate pending breakdowns, allowing practitioners to schedule maintenance or to design more reliable systems in the first instance. This is an area where the opportunity is clear but little work exists.

III. The state of fusion data

As they pursue these AI opportunities, scientists will need access to data from experiments, simulations, and sources that are not traditionally available, such as manuals and researcher notebooks. Scientists need access to much more of this data, but they also need the data to be as high-quality, or FAIR —findable, accessible, interoperable and reusable—as possible.

As a community, fusion researchers have not historically prioritised data curation, standardisation, and access, to the same extent as fields such as astronomy and genomics that have strong, often centralised, data infrastructures. There are promising efforts underway to address this, but also many obstacles.



1. **EXPERIMENTAL DATA: Unvalidated, single-machine and hard to access**
Experimental data is the 'ground truth' that the sensors in reactors pick up. It includes everything from line graphs to video. It looks different by machine, both between tokamaks and stellarators, and between different tokamaks. The data that a reactor generates also changes over the course of its lifetime, as the machine gets damaged. There is a continual need to extrapolate from data produced by today's machines to future reactors not yet constructed.

In magnetic confinement fusion, the challenge is not so much a *lack* of data, but an *excess* of *raw* data. As outlined in a recent paper by researchers at Oak Ridge National Laboratory, this raw data should ideally undergo several steps to make it useful to AI training. In the pre-processing phase, it should be *cleaned*, *calibrated* and *validated*, to address missing values, noise, and to account for imperfections in the underlying diagnostics —a task that requires expert engineers. In the post-processing phase, the data should be *labelled* and *annotated,* to capture 'features' that are important to training an AI model, such as the rate of change in the electrical current. The data should then be *organised* into formats that are suitable for the kinds of parallel processing that AI training uses.

Today, only a small fraction of fusion data goes through this kind of end-to-end processing and analysis. As a result, the community has to rely on the small number of well-validated datasets that are available, which one interviewee estimated at just a few hundred or thousand experimental 'shots'— the term used for individual test runs of a reactor.

The high cost of fusion experiments also creates a natural incentive not to fail. This caution curtails more novel experiments and means that many have fallen in a similar 'parameter

space'. This means there is not enough validated data, and the data that does exist is not representative of the full range of plasma dynamics that scientists want to model.

The experimental data that does exist is also not generally available open source. For example, the 40-year JET experiment generated more than 100,000 experimental shots. But interviewees estimated that only 5-10% has been properly validated, annotated and analysed. EUROfusion, the pan-European network of fusion research labs that managed JET, grants access to this data upon request, but typically only to researchers whose institutions have pre-existing EU agreements, and for research rather than commercial use.

Happily, practitioners are pioneering new approaches for the AI era. The UKAEA recently managed to successfully open source experimental data from their MAST facility. Several interviewees pointed to this FAIR MAST project as a best practice for the field. Some companies and students already use FAIR MAST, although data from the follow-up MAST-U facility is still embargoed (see below) and the platform is hard for non-experts to navigate.

Most AI fusion researchers also want *multi-machine* data, aggregated across reactors, rather than data from a single facility like MAST. This would allow them to cover a richer parameter space and demonstrate that their models are general enough to work across different machines, including those not yet constructed. In that vein, the IAEA is developing a Fusion Data Lake, with partners from across the world, including UKAEA. Given the scale of fusion data, the goal is not to create a single database, but rather a *federated* initiative, where institutions would store data locally but researchers could access it via a central *data catalog.*

The biggest "time sink" for such efforts is harmonising data from different machines, as facilities have traditionally followed different regimes. Some define and measure important phenomena—such as plasma disruptions—differently, while others use different 'structures' or schemas to index and link data points. ITER's Integrated Modelling & Analysis Suite, or IMAS, aims to address these issues by providing a standardised ontology and structures, as well as an API that scientists can use to analyse data that conforms to this standard.

Scientists have successfully retrofitted some older experimental data, for example from JET and MAST, to conform with IMAS and more facilities are planning to do so. However, interviewees cautioned that IMAS is not a definitive solution to standardising fusion data for AI. Retrofitting older data typically requires extensive inputs from the experts involved in creating that data. Practitioners also designed IMAS for standardising data from individual experiments, not for training AI models on data from many experiments. IMAS also does not yet cover all the physical quantities or abnormal events that scientists are interested in.

2. **Simulation data: No clear incentives, process or place to host it**
When researchers run simulation codes, the resulting data, for example on the rate at which heat leaks out due to plasma turbulence, could be used to train AI surrogate models to predict these variables at a fraction of the computational cost.

The challenge is that scientists are not currently incentivised to create such datasets and cannot simply package up old simulation data and turn it into a new database. Most fusion scientists run a simulation to answer a single, narrow, physics question. They do not run a large number of simulations to build representative datasets to train AI surrogates—a very different activity.

Most simulation codes contain 'free parameters'—knobs that scientists must decide how to best tune—a practice that can be as much an art as a science. To create a rich training set for an AI surrogate, scientists need to run the simulations at scale in a way that provides representative coverage of these parameters. There is no standard procedure for this work, and the codes are often finicky, requiring significant expertise to run. The resulting datasets can be huge and there is no obvious location to store them. There is also no equivalent to FAIR-MAST, UKAEA's open source platform for experimental data, although promising early examples exist.

Similar to experimental data, practitioners also want *multi-code* databases that harmonise data from many simulation codes, rather than relying solely on a single code. However, most simulation codes do not solve the same problem, which leads to inconsistency issues across them. As one interviewee noted: "*We have workshops where the best people in the world, using the best codes, are given the same simple problem... and the results are all over the place.*" Despite these challenges, progress is underway, with some codes already integrated with the IMAS fusion data standardisation system.

3. **Dark data: Nascent, IP issues, and hard to integrate into workflows**
'Dark data' describes the contextual information that scientists generate that could be very useful to training AI, but is not captured in structured datasets. It may exist in old papers (published or unpublished), presentation slides, or in the heads of a small number of people.

For experiments, dark data includes the free text notes in physicists' logbooks, where they describe the procedures they ran, the hardware issues they faced, or the phenomena they observed. It includes the vast amounts of information that fusion institutions maintain about their machines' design and diagnostics which is often scattered across web and file servers, with low visibility. Even published fusion journal articles may contain a lot of unstructured data.

For simulations, dark data includes the knowledge needed to run a code and interpret its results, including how to account for known but undocumented imperfections. For example, a code's documentation might fail to mention a major oversimplification that, if unaddressed, would lead to a dramatic overestimate for a variable of interest, like energy loss. Other codes may have legacy "do not touch" modules that were added by a code's creator, are necessary to a successful simulation, but which even the code's leading users do not fully understand.

Ultimately, this dark data often captures tacit knowledge that is not present in traditional databases.If fusion scientists and those training AI models could access this dark data, it could help ensure that they do not focus on the wrong things, for example when an anomaly in the data is caused by an accident or equipment failure, rather than a meaningful phenomenon. It could also provide AI models with a window into the *entire* research process, including its many

dead-ends, rather than just the final result. Once experiments like JET are decommissioned, or a code designer retires, this knowledge risks being lost forever.

Researchers are now using LLMs to try to make fusion dark data accessible to scientists. Early efforts include using LLMs to query experimental logs, archive documents, and to build a knowledge graph of important fusion concepts. However, these efforts face challenges. Much of the data is not sufficiently annotated to be useful. Researchers looking to extract data from logbooks or journal papers often face IP issues. It's also not yet clear how to reliably integrate this dark data into practitioners' daily workflows.

The three 'debts' holding fusion data back

Many of the challenges with accessing high-quality experimental, simulation and dark fusion data are symptoms that result from three underlying issues–technical, bureaucratic and cultural–which have compounded over time into systemic 'debts' that inhibit the use of AI today.

1. Technical debt

In the fusion community, the priority has been getting large machines to work, rather than building supporting infrastructure to generate, process, store and share data. As a result, activities like data curation and producing high-quality code have long been underfunded.

This underinvestment has compounded over time and now poses clear obstacles to using AI. For simulations, many of the leading computer codes were created decades ago and have evolved slowly, making them ill-suited to modern GPU-style hardware. For experiments, a lack of investment in new diagnostic equipment has led to both a lack of measurements and low-resolution measurements. Experiments also lack robust systems to process the data that they generate and to document its provenance. Exceptions exist, as with the robust data provenance systems at JET, but this expertise risks being lost as the systems are decommissioned.

2. Bureaucratic debt

Fusion's traditional reliance on government funding and international collaborations has led to obstacles that impede data sharing, including:

- **Complex ownership:** Many fusion projects have a web of political and contractual links across many countries and institutions. This makes collaborations, like multi-machine databases, difficult. For example, JET was sponsored and funded by Euratom, the EU's nuclear research community. EUROFusion, a pan-European network of fusion research labs, managed its scientific exploitation. UKAEA managed its engineering and operations. Releasing its data may require agreement from all actors.

- **Embargo periods:** Scientists that run fusion experiments often want an embargo period on the resulting data so that they can prepare a publication. Such embargos are rational, common in science, and largely supported, up to a point, but many interviewees felt that they have become too long.

- **Open-source policies:** Different government bodies have different policies about whether to open source data from the fusion research that they fund. For example, the MAST experiment was funded by UK Research and Innovation, which has strong open data requirements. The follow-up MAST-U experiment is funded by the UK Department for Energy Security and Net Zero, which does not have the same policy. Many private fusion companies do not open source their data at all.

- **Intellectual property:** IP poses obstacles to sharing data, and may become more of a challenge as commercially-viable fusion becomes more plausible. STEP Fusion, the UK's project to put fusion electricity on the grid, illustrates the tension. A newly created company that is a subsidiary of UKAEA leads the initiative. If STEP succeeds, using decades of publicly-funded research, how should the UK public benefit? One argument is that open-sourcing any accompanying data may result in the loss of a sovereign asset and primarily benefit large multinational companies. A counterargument is that fusion is a global challenge, the UK benefits from others' open data, and open-sourcing will enable the UK to attract talent and funds. A separate IP question relates to fusion companies: how should they share data from their new experimental facilities?

3. Human and cultural debt

Fusion lacks software engineers and experts who are able to clean data, attach confidence levels, and curate it for AI use. As one interviewee noted: *"There is no established career path for people to build and maintain databases. Twenty-five years ago, someone told me that databases were an activity of 'low intellectual level' with 'no prospects'."*

As a result, physicists must take on many tasks that are outside their core areas of expertise, such as writing high-quality code. On one hand, interviewees argued that the UKAEA's open-sourcing of data from the MAST experiment was only possible because a small number of dedicated data engineers took it on as a passion project — *"If you left it to the physicists, they would tinker forever."* On the other hand, to validate and interpret the data, data engineers need the expertise of the physicists and engineers who designed the original simulations and experiments.

The skills issue is compounded by a research culture that inhibits data sharing. Interviewees stressed that scientists are always pushed to move on to the next experimental campaign, rather than going back to validate old data. As a result, a lot of fusion data is incomplete. This can stop scientists from sharing their data, because they fear that end users will not appreciate the nuances and do bad science with it. Or they fear that they (the experts) will be criticised as "unscientific" for releasing the data. Some interviewees argued that the UKAEA was only able to open source data from its MAST project, because many of the scientists who ran the original experiments had moved on to the follow-up MAST-U project.

IV. Recommendations

A future vision for UK fusion, with respect to data and AI, would look something like as follows:

- **Experimental data**: All useful data from JET and MAST-U is openly available. The UK's leading fusion facilities have strong end-to-end pipelines for gathering and analysing new experimental data. A significant portion of this data is validated and annotated, in line with emerging standards, like IMAS. Researchers are able to access this data alongside data from other machines, in shared databases or federated systems, with clear protocols for how to use and contribute to it.
- **Simulation data:** The most important simulation codes are modernised, optimised, documented and maintained. Researchers are explicitly funded to generate large-scale, representative data from these codes, and to train AI surrogate models on this data, to improve the overall simulation process. For key phenomena, researchers have access to multi-code databases that do not rely on a single simulation code.
- **Dark data:** The tacit knowledge involved in designing and running experiments and simulations is captured and made queryable to the fusion community, with explicit funding and tooling.
- **Skills and infrastructure:** Data and software engineers are embedded within fusion research groups. Data infrastructure is recognised as pivotal to future progress.
- **Governance:** There are clear and permissive policies to access and use fusion data, including in commercial settings, with a time limit on any embargoes.

How to get there? We provide **eight recommendations**.

- We focus on projects that could be funded and implemented within 1-2 years. But each project should also provide insights for how to address longer-term issues impeding fusion data.
- Each project could be led by a combination of government research organisations, like UKAEA; government departments and research funders, like the Department for Science, Innovation and Technology and UK Research and Innovation; AI and fusion companies; universities; and philanthropic funders like Renaissance Philanthropy and Google.org. The UK should also look to collaborate internationally, for example with the US Genesis Mission and IAEA.

1. Strengthen the UK's lead in open fusion data

- Expand FAIR MAST, the UK's pioneering open-sourcing of experimental data from its MAST facility, by adding data from the follow-up MAST-U facility and making the user interface more accessible.
- This will require the UK Department for Energy Security and Net Zero clarifying that open data policies apply to MAST-U, funding at least five data engineers over a two-year time period, and ensuring that the project has sustainable compute and data storage. A key question is how to best host and maintain the database over time.

2. Liberate 40 years of data from the Joint European Torus

- Launch a project to open source at least 30% of JET experimental data by 2028. This will require addressing two major questions, of which the second is the hardest.

    - First, what specific data to release? Should the project focus on the subset of JET data that is already well-structured, validated, and curated? Or should it also include raw data, or data that is only validated in part, for example to account for errors in the instruments? Should it only release 'important' scientific data, such as that associated with notable discoveries, or should it also include data from experiments that were deemed 'uninteresting' at the time, but which may offer valuable insights into 'normal' machine behaviour for AI systems?
    - Second, how to most expediently solve the political, bureaucratic and legal challenges of securing agreement between all the institutions originally involved in JET, in order to release the data?

3. Launch a competition to predict plasma disruptions

- Fund a competition to see which AI model can best predict the plasma disruptions that are the primary obstacle to a working fusion power plant. This should build on other AI fusion competitions and disruption prediction efforts, for example by the Disruption Group at MIT and the US Department of Energy. It should also focus on predicting disruptions in new experimental campaigns, rather than historical data. This may require dedicated funding for new experimental shots on machines such as MAST-U, so that models can be tested on challenging edge cases.
- The competition designers should also specify the criteria that they want to evaluate models against, in a way that incentivises the most useful and robust models. In addition to accuracy in predicting disruptions, key criteria, and sub-competitions, could include:
    - Data efficiency: Is a model able to predict disruptions when it only has access to small amounts of data? For example, if the data made available to a model is reduced to a smaller set of diagnostics, to better approximate a scenario where some diagnostics in a reactor have become damaged, does it still perform well?
    - Transferability**:** Can the model predict disruptions across different reactors?
    - Lead time: Can the model predict disruptions with sufficient lead time to prevent them?
    - Interpretability**:** Can the model shed light on why disruptions are occurring?

4. Prototype the future of AI-enabled scientific data curation

- Expand the platform that UKAEA recently developed to enable human experts to use AI to annotate data from the MAST experiment, for example to identify and add details on plasma disruptions. This project could make the platform more useful in several ways:
    - Adding data from other fusion facilities;
    - Increasing the complexity and variety of the metadata that is captured;
    - Training AI models to directly annotate an increasing share of this historical data
- Done well, this project could help establish the UK as a leader in AI-enabled scientific data curation and serve as a guide for other disciplines for how to best combine human expertise and AI to annotate scientific data at scale.

5. Make leading simulation codes AI-ready

- Launch an effort to modernise the most important fusion simulation codes. This will enable researchers to run these codes more efficiently, build up large representative datasets, and train AI surrogate models on this data to improve simulations. This effort should build on recent work to develop AI-ready fusion codes, such as GSFIT from Tokamak Energy, and FreeGSNKE by UKAEA and STFC researchers, as well as community efforts to improve plasma research software like PlasmaFair.
- This project will require first aligning on priority code(s) to modernise:
    - One potential target is JINTRAC, a suite of codes that are important to the UK's proposed STEP power plant and the international ITER effort. Tools like Torax have made progress in modernising modules within JINTRAC, but others, like the modules that researchers use to simulate the "scrape-off" layer where the exhaust heat hits a reactor's walls, are slow and cumbersome to run. The modules focussed on how the plasma is heated, how heat and particles leak, and how to inject fresh hydrogen pellets, would also benefit from modernisation.
- Once the priority codes are identified, the project could modernise their documentation, including updating their configuration, user manuals and user interfaces. In a second phase, the codes could be 'refactored' so that they are compatible with modern chips like GPUs and for massively parallel data generation. The project could also test the usefulness of LLM-based coding tools in this modernisation process. After this, the project could open source the codes and outline a plan to maintain them, longer-term.

6. Demonstrate a new state-of-the-art for AI surrogate models

- Fund small teams of software engineers and fusion experts to develop AI surrogate models for important computationally-expensive processes in fusion simulations, such as transport, heating, and fuelling. These AI surrogates would provide the fast predictions researchers need to simulate the entire lifespan of a fusion plasma, across many scenarios.
- To produce AI surrogates that are usable in production, and don't quickly become out-of-date, the project should ensure that all new surrogates have state-of-the-art documentation, data provenance and version control. It should also release the data used to train and validate the surrogates. To drive the biggest long-term benefit, the project should develop libraries and software pipelines to automate time-intensive aspects, such as organising the data, building on early examples.

7. Use AI agents to preserve expert fusion knowledge for the future

- Launch a project to make the tacit knowledge involved in running leading fusion simulation codes accessible to the wider research community, led by experts, such as the code's developers and/or its leading maintainers and power users,
- To do so, use an AI coding agent to run the simulation code. As it seeks to execute the code, the agent would have an "internal monologue" that is traceable. The experts would oversee the process, guide the agent, and be able to intervene and steer it. The end result would be a series of documents, such as markdown files, that capture the "dark data" for how to run leading fusion codes. This would extend far beyond the API needed

to run the code, and capture the “hidden knobs” and guidance that are needed to run leading fusion codes well, but which only a small number of human experts understand.

8. Create Fusion-Bench to measure and drive LLM performance

- Task leading fusion experts with creating a new evaluation method to understand how well leading large language models understand core fusion concepts. Currently, there are no strong fusion benchmarks for LLMs. Addressing this would make it easier to evaluate and improve the usefulness of the models for downstream tasks in fusion.
    - Such an evaluation would be more difficult to create than in other disciplines like maths or computer science, because it will be hard to automatically verify a model’s performance. But, experts could use this project to align on the most useful approach, which may involve a combination of question-answering and carrying out various types of tasks.

V. Six open debates

Interviewees disagreed on some points. Here are six debates worth noting. Despite the framing, few are simple either/or situations. Rather, most debates are about relative degrees of emphasis for funders of AI fusion data projects.

- **Incrementalism vs Novelty:** Should we build on the early AI opportunities that fusion practitioners have already showcased or pursue more novel, uncertain, ideas? In our recommendations, we generally went with the former, but some experts suggested focusing more on the latter. More novel ideas could include building datasets to train general-purpose ‘fusion foundation models’ to help with everything from disruption prediction to agent workflows. Another idea was using AI ‘world models’ to pursue new kinds of fusion simulations that don’t require scientists to make as many assumptions about the underlying physics involved.
- **The Past vs the Future:** Should we try to get as much value as possible out of older fusion data, like data from JET, as it remains the closest the world has ever come to a functioning power plant? Or, do the costs involved in cleaning, validating and annotating old experimental data mean that we should instead accept our losses and focus on making future fusion experiments AI-ready?
- **Science vs Engineering:** Are efforts to validate, annotate and standardize fusion data, like IMAS, part of an ultimately doomed quest for perfect “human scientific understanding” in fusion? Should we instead embrace a more “engineering-led” approach that can work with imperfect, noisy data? One version of this debate is whether it is better to release imperfect data early, with caveats, or to only release high-quality verified data.
- **Domestic vs International:** Should the UK rejoin ITER, the world’s flagship international fusion collaboration, which the UK left following Brexit? Or should the UK focus on advancing its domestic efforts, like MAST-U and STEP, perhaps in collaboration with a small number of priority partners, like the US?

- **Magnetic vs Alternatives:** Should the UK continue to focus disproportionately on magnetic confinement fusion, as the most realistic pathway to a future power plant? Is magnetic a better bet for AI because it produces much more data and doesn't have the same associations with the security establishment, which makes data access for magnetic easier? Or should the UK invest more in inertial fusion and hybrid alternatives, given the country's academic expertise, its historically strong relationship with the US National Ignition Facility, and some notable assets, such as the UK Central Laser Facility's world-leading laser.
- **Public vs Private:** Should the UK government try to derive more direct immediate value from its fusion investments and data? For example, should the UK license some public data at cost to companies, to help cover the costs of data processing and annotation? If so, should local startups pay less? Or would such efforts hurt the UK's goal of having a world-leading fusion sector?