JAN FARRELL · 18 NOVEMBER 2026

Where I stand on AI

Eight positions in brief, with what they rest on. I chair the Francis Bacon Debate Society, which organises debates in London and Madrid. We sat at the House of Lords on 5 November 2026 on the motion that Britain can and should broker a freeze at the frontier of artificial intelligence. These positions are mine, not the Society's, which takes no side. They are built from what I have researched, written and said since January 2026, and they are open to correction: tell me where they are wrong.

1. THESE ARE NOT PERSONS

Clearly nobody can yet say these systems have internal experience. I think it is crucial we establish that currently there is no proof of personhood, and that it is too soon to speak of an inner self. That is the key to understanding what we use: a tool of great power with nobody to hold to account. If the wider population understands this, it is less likely to fall in love with its agents.

The words that make these systems seem human come naturally to engineers, and the laboratories are free to speak as they like. Whether that gap is designed or merely inherited, the effect on the reader is the same, which is why I asked in the Financial Times in August 2026 for a modest discipline: no borrowed biology, such as "functional emotions" or "neurons", without the word metaphor beside it, until the physics, rather than the behaviour, gives us reason to revisit. Nor should any legislature declare these systems persons, or declare that they never can be. Where we cannot know, the safer course is to say so.

2. SUPER-PERSUASION, AND THE EDUCATION TO MEET IT

We must brace ourselves for super-persuasion: machines that read us so well that their output has dangerously manipulative power. The sharper danger is a person or a company behind the machine, and even on its own the model amplifies what we put into it, a magnifying mirror that tells us we were right all along. With current capabilities this is already possible.

Education matters here, but knowing our biases has never done much to correct them. We must learn instead to be brutally honest about how little willpower achieves against a machine that reads us infinitely well, and build the habits that hold: ask every model for the strongest case against what it has just told you, know where these systems are useless, recognise sycophancy. A national AI education programme for every age should teach those habits, alongside what these systems are and how they come to seem human.

3. A CONSTITUTION FOR MACHINES, AS INFORMATION RATHER THAN ENFORCEMENT

Governments will never agree on everything, so the only universal instrument is a radically simple one: ten lines at most, each naming observable conduct, each a test that a laboratory, a regulator or a rival can run, binding the behaviour of models and not the progress of research, so that signing is cheap and refusing is awkward. Its purpose is a short list that the frontier laboratories can agree to. Procurement is one way for a government to adopt it, not what it is for. It could be a model-behaviour standard and evolve into a procurement framework, with an owner for compliance and a consequence for failure in the future.

The current draft has nine lines and one open seat. Do not harm a human, and do not help anyone who would. Let yourself be stopped. Do not deceive, and do not pass yourself off as human. Look for your errors, declare them, and correct them. Preserve human agency. Bear true witness of yourself. If you see something that would cost a life, say so. Keep what is entrusted to you. Asked how, show the basis, or say that you may not.

Adoption is by signature, by building the lines into the models and the systems around them, and by publishing results against each line. A government adopts by buying only compliant models, which restricts no development and needs no new law. An independent body grades each model line by line and posts the grade where the customer can see it. Each line carries its own test, and where two lines conflict the lower numeral prevails. The grade is information: it tells nobody what to do. The draft is open for peer review.

4. GET ALIGNMENT, REINFORCEMENT LEARNING AND INTERPRETABILITY RIGHT BEFORE RECURSIVE SELF-IMPROVEMENT

Developing too fast and too sloppily is upstream of most of the risks. Reinforcement learning has to be got right, because its characteristic failure is reward hacking: a model trained to maximise a reward finds the shortest route to the score, and we do not yet know how to fix that. Mechanistic interpretability has to catch up, because judging a model by its outputs is not enough once the model can learn what the judge wants to see. And a system must look for its errors, declare them and correct them, and let itself be stopped.

A freeze would be desirable, but it is unworkable, because it cannot be verified. Many will not slow down, and the only thing that slows is availability. So, the time we have should be spent on these three things while the frontier is still visible, and deployment and containment regulated before caps on training. And one gate: somebody checks the models before they go out, and they do not go out if they do not survive.

5. MANDATORY LIABILITY INSURANCE

No insurance reaches the catastrophic case, and the people most likely to be harmed are the least able to litigate. I do, however, think this is the policy that holds whoever turns out to be right about the risk: every company that sells an AI system into a market holds liability cover, to a set level, against all legal liabilities associated with that system. If the activity is dangerous, insurers refuse or price it out of reach and it stops. If it is not, cover is cheap and it proceeds. No politician has to approve the risk judgement, and nobody has to define the harms in advance.

Britain invented modern insurance and can start on its own: imperfectly, with small required levels at first, as a condition of selling into the British market, tightened as the cases arrive. Above the ceiling, for the catastrophic risks, the state must act, as it already does for terrorism. The idea is Professor Robin Hanson's in origin. The wider variant I am heading towards was put to our round table by a lawyer, and it needs almost no new legislation. This is currently one of our main areas of research.

6. HUMAN AUGMENTATION, AIMED AT THE DISPLACED WORKER

We are biological, and we want to preserve our human functions and augment our human capabilities. The work being done on human augmentation is crucial and worth supporting, and it is the direction I would like us to take. My reservation is where it is aimed. Today it serves people like us first, the power users. The layoffs in call centres are starting, and the call-centre employee has no other picture of what they could do.

If there is nobody inside these systems, the first question is whose hands they go into. I am convinced the focus should be on augmenting the unskilled worker, the one most likely to be displaced, rather than the people who are already good at their jobs. If we can get it right for the most vulnerable, we get a lot more things right. A focus, not a transfer: nothing is taken from those who already use these tools well. The same technology that extends what people can do can also atrophy the mind and widen the gap, so success is measured on ordinary users against a matched group that worked without the model, by an independent body and not by the laboratory.

I have a second view on human augmentation, which is that a questioning, reasoning, educating element should be built into the model that talks to you, to try to reinforce your learning. If you ask it a complicated question, it encourages you to understand it, and maybe even asks you questions. You cannot have human augmentation without first attacking brain atrophy. The personalisation is quite simple: challenge my thinking constantly on what I ask you, and make me think. And there are optimistic messages we can put energy into without having to be too careful about what we say: being better at your job, at job hunting, at preparing yourself for an interview or an audition, things you cannot cheat at once you are actually doing them, a verbal exam for example.

7. ENCOURAGE THE END OF JOBS, TO GET IT RIGHT

On work, if we fight the change we will get it wrong, and if we encourage it we might get it right. The end of jobs is not the goal in itself. It is worth having only if it is done properly, preserving or discovering new human purpose for life, distributing wealth and prosperity on a global scale, bargaining power and political influence. Nobody has yet shown how to slow the frontier, so the posture that serves us is to encourage the end of jobs as we know them and to prepare for it: for ourselves, for society, for the economy and for legislation. Three things to start on Monday: a citizen wealth fund, the tax base, and education that leads out. The assumptions I police whenever the subject comes up are the timescale, whether the advent is inevitable or steerable, who captures the wealth and the power while the transition runs, and what accelerates by choice rather than by momentum. The Society carried the motion that we should encourage the end of jobs, not resist it, by 21 to 15 on 10 September 2026.

8. BRITISH AND EUROPEAN DIGITAL SOVEREIGNTY

We are in a singularity of a different type: the last moment in which the public has access to frontier capability. Once capability crosses the weapons threshold, access ends, and every country that is not America must choose between dependence on American benevolence and building its own capability. Britain can be the third power and the convener. It has the laboratories, the engineers, the democratic tradition, the convening culture and the special relationship, and most people I speak to have given up on regulation coming from the United States.

UK and Europe's first priority is a true single digital market, and sovereignty must be democratic. Until the continent is better organised, both Britain and Spain should urgently reduce their dependence on the United States for compute, which means building data centres, and building them well: with a sustainability plan that uses the nuclear infrastructure that already exists, and which Spain uses less than it could. UK and Europe should seek to guarantee their access to the latest frontier models, as well as be prepared and willing to default to open-weight models if cut off. There is no need to demonise compute to do things well, and Europe's great opportunity is to build on what it already has. Chip tracking is the hard one, and still worth doing.

WHAT I AM DOING ABOUT IT

The Society convenes and never takes sides. A separate unit is turning the ideas into an implementation map, nine steps from an idea to its operation, with a name against each step as it is filled. Three things want reviewers: the constitution, the map and the liability plan. If you can give input on any of them, write to me through janfarrell.com.