Taming AI: Living with a New Kind of Power
Professor Matt Jones
Tuesday 21 April 2026

Over the course of this lecture series, we have explored a set of unsettling possibilities about artificial intelligence. We have imagined it as an overlord, wondered whether we might be assimilated into its logic, and reflected on the subtler risk of becoming domesticated by the comforts it offers. In the last lecture, “Born Supremacy”, we paused that trajectory and asked a more useful question: not what AI might become, but what we are. The answer that emerged was that human intelligence is not merely about performance, but about inhabitation. It is embodied, situated, and bound up with consequence and responsibility. This leads us to an important distinction between AI and ourselves, one that allows us to constructively think about how we might live alongside it.

If AI is not a being in the way we are, then what is it? My central claim in this lecture is simple but useful: AI is best understood not as an intelligence to rival us, but as a form of power. And like all transformative powers throughout history, it can – and must - be tamed.

Learning from Fire: Control, Not Elimination

Fire is one of the earliest and most world-changing powers humanity encountered. It enabled cooking, warmth, and industry, but it also brought destruction when left uncontrolled. The Great Fire of London in 1666 was not a failure of fire itself, but of the systems surrounding it. What followed was not the abandonment of fire, but the redesign of the world around it: new building materials, new regulations, and new forms of coordination.

This is the first lesson of taming a power. We should not knee-jerk like try to eliminate it. We should work to develop systems that allow us to live with it safely, predictably, and under human control. The question for AI, then, is not how to stop it, but how to construct and deploy it so that it fits within human capabilities and responsibilities.

The License to Control: Humans in the Loop

Consider another familiar power: the car. When cars were introduced, society did not simply release these powerful machines into the world. It built roads, rules, and, crucially, a licensing system. A driving licence is not just permission to use a machine; it is a declaration that a human being (many of us!) have the skill and judgement to remain in control of it.

Now, as cars become autonomous, the meaning of that licence begins to blur. If the system no longer depends on human skill or attention, then what does it mean to be “in control”? This is the subtle way in which power can slip. Not through dramatic takeover, but through the gradual attrition of human engagement. We move from being active operators to passive supervisors, and eventually to disengaged passengers.

The driver to autonomous car move is a useful metaphor when thinking more broadly about AI, as Matt Crawford in his book “Why We Drive” explores in more detail than we have time for in this lecture The risk that leads to us ending up out of control is not that systems become more capable, but that we become less attentive, less skilled, and less willing to question.

Two Kinds of Systems; Verification and Containment

Not all AI systems are the same. Some are tightly bound, with clear tasks and measurable outcomes. Medical imaging systems, for example, operate within strict constraints. They analyse specific types of data, provide probabilistic outputs, and remain embedded within a broader system of human expertise and regulation. These systems can be tested, verified, and certified. They represent what we might call a “tamed intelligence.”

In contrast, general-purpose AI systems such as large language models are much more open-ended. They can be applied to a wide range of tasks, often in unpredictable ways. There is no single ground truth against which their performance can be measured. As a result, we cannot tame them through verification alone. Instead, we rely on containment: filtering inputs and outputs, monitoring behaviour, and keeping humans involved in decision-making. We are with these tools moving from a world where intelligence is a power we can fully specify, to one where it is a capability we must continuously shape and manage.

Legibility: Can We Read the System?

At the heart of this challenge is the concept of legibility. If we cannot read a system, we cannot control or govern it. In the lecture we consider how legibility can be seen to operate at three levels: data, reasoning, and behaviour.

At the level of data, AI systems learn from vast datasets that reflect patterns in the world. These patterns include biases and assumptions. If the data is flawed, the system will reproduce those flaws. Making data legible means understanding what has gone into the system and how it shapes its outputs.

At the level of reasoning, we face the problem of explanation. When a system makes a decision, we want to know why. Techniques such as LIME and SHAP provide approximations of how models arrive at their outputs, but these are not perfect explanations. They offer insight, not full transparency.

At the level of behaviour, the challenge becomes even more complex. A system may appear reasonable and even provide plausible explanations yet still behave in ways we did not intend. This is where the notion of alignment becomes crucial.

Alignment and Its Limits: Learning What We Appear to Prefer

Modern AI systems are often shaped through processes such as Reinforcement Learning from Human Feedback (RLHF). In this approach, systems generate outputs, humans evaluate them, and a reward model learns what humans seem to prefer. Over time, the system is optimised to produce outputs that align with those preferences.

However, there is an important nuance here. The system does not learn our values directly. It learns a statistical approximation of what appears to be preferred. This “shadow” of human judgement can be misleading or incomplete.

In the lecture I use the analogy of a well-trained dog. A dog may learn to follow commands reliably, but it can still chase a squirrel. At that moment, training is no longer sufficient. What is needed are constraints that limit behaviour. Which leads us to “guardrails”.

Guardrails: Constraints, Not Intelligence

Guardrails are mechanisms that constrain what a system can do. They do not make the system more intelligent; they ensure that certain actions are limited. The analogy of nuclear control rods illustrates this point vividly. These rods are not there to guide the reactor’s decisions, but to physically limit its behaviour.

Yet guardrails can fail, especially when they are poorly designed or interact with the system in unexpected ways. The Chernobyl disaster is a stark reminder that constraints must be carefully engineered. When they fail, they can do so catastrophically.

Similarly, attempts to define simple rules for AI behaviour, such as Asimov’s Laws of Robotics, often break down in practice. Asimov’s own stories demonstrate how conflicting rules can lead to paralysis or unintended outcomes. A robot following its rules perfectly may still behave in ways that are useless or even harmful.

Trust: Designing the Signals We Rely On

If full legibility is unattainable, then we must turn to trust. Trust is not an abstract concept; it is built from signals such as consistency, transparency, competence, boundaries, and accountability. We rely on these cues to navigate complex systems.

As we will see in the lecture, current AI systems provide these signals unevenly. They may appear confident but be wrong, leading to over-trust, or they may be dismissed entirely, leading to under-use. Designing trustworthy systems means making these signals explicit and reliable.

Beyond the Interface: Participatory AI

Much of what we consider up to this point in the lecture focuses on the point of interaction: adding explanations, inserting humans into the loop, and designing guardrails. While these are important, they operate at the surface. By the time we encounter a system, many crucial decisions have already been made.

Participatory AI shifts the focus upstream. It asks who defines the problem, who selects the data, and who determines what counts as success. These decisions shape the system’s behaviour long before it is deployed. If those affected by the system are not involved in its design, the system may fail to reflect the contexts in which it operates. It will be a power that does not reflect the inhabited, messiness of the people and world it is meant to serve.

Conclusion: Remaining Capable in Relation to Power

The lecture closes by returning to the central theme: taming a power is not about eliminating it, but about ensuring that we remain capable in relation to it. The danger with AI is not that it becomes too intelligent, but that we quietly withdraw our attention and responsibility. AI won’t take control from us; rather we will give it untamed control when we stop noticing. The question, then, is not whether AI will become more capable. It will. The question is what we

will stop doing as it does. Will we stop checking, questioning, and understanding? Or will we design systems, practices, and institutions that keep us engaged?

Because in the end, the story, as with all the lectures in this series, is not about AI. It is about us.

References and Further Reading

Legibility I — Data

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723
A practical proposal for documenting datasets, making visible how data is collected, shaped, and constrained.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown.
A compelling critique of how data-driven systems can encode bias and produce systemic harm at scale.

Legibility II — Reasoning

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You? Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144. https://doi.org/10.1145/2939672.2939778
Introduces LIME, a widely used technique for making individual predictions interpretable.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.
Presents SHAP, a principled framework for attributing model outputs to input features.

Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans.* New York: Farrar, Straus and Giroux.
A clear and grounded account of how AI systems "reason," and where that reasoning breaks down.

Legibility III — Objectives

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* New York: Viking.

A foundational text on alignment, reframing AI design around uncertain and evolving human preferences.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. “Concrete Problems in AI Safety.” *arXiv preprint* arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565
A practical overview of key challenges in specifying and maintaining safe objectives in AI systems.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. “The Off-Switch Game.” *arXiv preprint* arXiv:1611.08219. https://doi.org/10.48550/arXiv.1611.08219
A formal treatment of how to design systems that remain corrigible and responsive to human intervention.

Legibility IV — Behaviour

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, et al. 2021. “On the Opportunities and Risks of Foundation Models.” *arXiv preprint* arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258
A comprehensive account of large-scale AI systems, including emergent behaviours and systemic risks.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, et al. 2022. “Ethical and Social Risks of Harm from Language Models.” *arXiv preprint* arXiv:2112.04359. https://doi.org/10.48550/arXiv.2112.04359
A structured taxonomy of how modern AI systems can fail in practice.

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values.* New York: W. W. Norton & Company.
An accessible and richly reported exploration of how AI systems learn and mislearn human preferences.

Legibility V — Human Judgement and Practice

Crawford, Matthew B. 2015. *The World Beyond Your Head: On Becoming an Individual in an Age of Distraction.* New York: Farrar, Straus and Giroux.
A powerful account of attention, agency, and skilled engagement, offering a vital counterpoint to automation.

Crawford, Matthew B. 2020. *Why We Drive: Toward a Philosophy of the Open Road.* New York: Harper.
A rich exploration of driving as a practice of embodied skill, responsibility, and freedom, illuminating what it means for humans to remain meaningfully “in control” of powerful systems.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.
A foundational text on human judgement, bias, and decision-making, essential for understanding automation bias.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” In *Proceedings of Innovations in Theoretical Computer Science (ITCS 2017)*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
Shows that some fairness tensions in algorithmic systems cannot be eliminated, only managed.

Governing AI in Society

European Commission. 2024. *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act)*.
The EU’s risk-based regulatory framework for AI, offering a real-world example of how societies attempt to govern these systems.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, et al. 2018. “AI4People—An Ethical Framework for a Good AI Society.” *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5
An accessible overview of ethical principles for AI governance.

Coeckelbergh, Mark. 2020. *AI Ethics.* Cambridge, MA: MIT Press.
A concise introduction to the philosophical and societal questions raised by AI.