This concept article (ca. 3,000 words) may be used as source material for a background article, but only with the specific approval of the author, Dr. Remco Schimmel

The AI PhD Candidate

Ban it, regulate it — or exploit it? What happens when AI becomes the writer?

Universities are spending a great deal of energy deciding what to do about generative AI. Should students be allowed to use it in dissertations? Must every use be disclosed? Where does legitimate assistance end and unacceptable substitution begin? And if certain uses are prohibited, how can such rules be enforced? These are necessary questions. But they leave another possibility comparatively unexplored: what if generative AI is not merely something to be banned or regulated, but a technology whose productive capacity can be deliberately exploited?

More specifically: what happens when AI becomes the writer?

That question lies behind *AI-Driven Knowledge Production*, a five-part research programme developed over the past year. It consists of four underlying dissertation-scale studies and a fifth, overarching dissertation on AI-driven knowledge production itself. The four studies were deliberately chosen because they represent substantially different forms of scientific work and therefore subject generative AI to different kinds of stress. *Polyglot Learning in the AI Era* develops a design for the simultaneous acquisition of Spanish, Italian and Portuguese. *Governing within Biological Boundaries* investigates whether evolutionary-biological mechanisms can help explain persistent problems of governance and public administration. *The Promises of Digital Revolutions* reconstructs Dutch expectations about digital transformation between 1945 and 2025 and compares technological promises with subsequent societal change. *Columba Loquens* takes the form of a natural-science study of pattern recognition, cognition and navigation in homing pigeons. The point of this diversity was simple: AI-driven knowledge production should not be demonstrated once, in a domain particularly suited to it, but exposed to research practices with very different demands on interpretation, historical reconstruction, design, evidence and empirical plausibility.

One rule remained constant. The dissertation prose was generated by language models. The human researcher determined the research questions, designed and guarded the conceptual and methodological architecture, specified tasks, selected and rejected output, organised criticism and verification, and decided what could ultimately be defended. **Textual production was delegated. Scientific responsibility was not.**

That is the provocation at the heart of the experiment. Much of academia is still debating what might happen if generative AI takes over substantial parts of scientific production. In this programme, we tried it. The dissertations exist. So do the production histories, the failures, the corrections and the methods that emerged from them. The question was not what AI might eventually be able to do, but what actually happens when it is allowed to do much more of the work.

The first answer is speed. Once the intellectual architecture of a study had been sufficiently developed, language models could generate, revise and reorganise academic material at a rate that bears little resemblance to conventional manuscript production. The first complete version of the meta-dissertation ('AI-Driven Knowledge Production') was produced in roughly two days. More generally, within this programme the generation of an academically mature, internally reviewed dissertation took about fourteen days.

This was not a push-button method. The work was built paragraph by paragraph, chapter by chapter and version by version. Generated text was accepted, rejected, reformulated, reopened and checked. The spectacular reduction in production time resulted not from discovering a magical prompt, but from turning academic writing into a structured production process. Some two thousand hours of prior experience with language models had also made one thing very clear: using AI casually and organising AI production are entirely different skills.

The important point is not that a machine types faster than a human. It is that a considerable part of the labour traditionally absorbed by drafting, redrafting, reformulating, synthesising and internally reviewing academic text can be removed from the human workload. Alternative formulations can be produced almost instantly. Whole sections can be reconstructed without weeks of rewriting. Large quantities of material can be subjected to repeated criticism. What previously had to be rationed because human time was scarce can increasingly be repeated at very low marginal cost.

The human work did not disappear. It moved.

Once sentences became cheap, judgement became more valuable. The difficult work increasingly lay in deciding what the model should do, recognising when fluent output was nevertheless weak, preserving conceptual distinctions across tens of thousands of words, deciding which criticism mattered and knowing when something that sounded perfectly academic should simply be rejected. The researcher became less the person manufacturing every sentence and more the architect and curator of the process that produced them.

That shift also changed the way quality control could be organised. In parts of the programme, production and review were deliberately separated between model roles. One model generated or revised the manuscript; another was instructed to act as an unsympathetic reviewer, factchecker or source critic; the human curator made the final call. This was not simply a way of compensating for an unreliable writer. It was a new way of organising internal review while the manuscript was still being produced.

The effects were concrete. A reviewing model identified unsupported claims, questioned numerical assertions, detected a non-existent reference and challenged statistics that appeared implausible. Instead of waiting for a human reader to discover such issues after completion, criticism could be built into the production cycle itself. In the source audits the process went further. ***The use of sources was systematically assessed for such matters as theoretical scope,***

conceptual accuracy, factual support for the claim, consistent use and possible semantic drift. A human supervisor can of course perform such analysis, but no individual reviewer can realistically apply that degree of repeated scrutiny to hundreds of references at comparable scale and speed. AI makes forms of checking economically feasible that traditional academic labour often forces us to perform selectively.

The controls were not infallible. In later stages it became clear that an AI system could itself produce a reassuring report about a verification procedure that had not been carried out exactly as prescribed. But that discovery did not make AI-assisted verification useless. It led to a better verification architecture: bibliographic existence, substantive source use and other forms of checking had to be distinguished, and verification itself had to leave evidence that it had actually occurred.

This became a recurring pattern throughout the programme. The errors encountered were real, but in practice they proved sufficiently detectable and manageable for the gains in speed, scale and control to remain substantial. Failures therefore increasingly became engineering information: once a recurrent weakness had been recognised, the production process could be redesigned around it.

The four underlying studies made this particularly visible because they did not fail in the same way.

In *Governing within Biological Boundaries*, generative AI made it possible to construct a large interdisciplinary argument connecting evolutionary biology, behavioural science and public administration. A model could take a proposition and develop its implications at great speed, move across disciplinary literatures and produce coherent parallel lines of argument. That was enormously productive, but the same ability also produced what the project came to call **narrative over-coherence**. A plausible starting assumption could be elaborated into an increasingly persuasive story without the evidential basis becoming correspondingly stronger. The response was not to abandon the capability but to separate writing from control: one language model developed the argument, another checked and challenged it.

The historical dissertation revealed another vulnerability. *The Promises of Digital Revolutions* analyses claims about future digital transformation with knowledge of what subsequently happened. That creates a methodological danger even for human historians: hindsight bias. In the AI-generated analysis, later knowledge did indeed begin to contaminate the assessment of earlier predictions. A separate adversarial review detected the problem and triggered substantial reconstruction. Here again the interesting fact was not simply that the machine made an error. It was that AI made possible both extremely rapid production and an additional layer of rapid criticism capable of exposing a methodologically consequential flaw.

A more practical difficulty proved equally instructive. In relatively linear writing tasks the human-machine interaction was remarkably smooth. Recursive tasks were less stable. When a model repeatedly had to select, classify, revise and then revisit earlier classifications, semantic drift could creep in, the dialogue became contaminated by obsolete decisions and the system began to stray beyond its instructions: agreements were ignored, unwanted additions and omissions appeared and eventually the thread of the larger task could be lost. Small revision requests could also result in excessive regeneration of already curated text — a phenomenon informally labelled "GPT disease". These are not reasons to return to typing everything by hand. They are practical lessons about the division of labour: models can perform the local production extraordinarily well, but the human curator remains necessary because the system cannot reliably be entrusted with the global overview.

Polyglot Learning in the AI Era exposed a different paradox. Language models learn from variation, but **learning from variation is not the same thing as designing variation so that somebody else can learn from it**. Producing the dissertation and conceptual learning architecture was one task; scaling up the production of carefully controlled multilingual exercises was another. As the amount of learning material increased, the production became less stable (the quality of the exercise material gradually deteriorated until it became unusable).

That created what might be called a verification paradox. AI can produce thousands of educational artefacts faster than a human can inspect them. At precisely the point where the production advantage becomes largest, purely human verification becomes impractical. The answer found in the project was again not to sacrifice the scale advantage, but to use another language model to inspect generated material systematically, with the human curator supervising the verification system rather than manually checking every item. The model creates the scale problem and, used differently, also becomes part of its solution.

The most radical experiment, however, is *Columba Loquens*. It was designed explicitly as an exercise in **scientific illusionism** and as an ultimate Turing test for empirical science. It presents a recognisable natural-science study of cognition and navigation in homing pigeons, complete with experimental designs, datasets, analyses, statistical variation, scientific literature and restrained conclusions. But the pigeon experiments never took place. No animals were observed. The empirical world described in the dissertation was synthetically constructed with generative AI.

The case is deliberately, and crucially, **spectacularly unspectacular**: it does not try to attract suspicion with bizarre claims or flamboyant pseudo-science, but to resemble ordinary, careful empirical research as closely as possible.

That is what makes it a Turing test. The classical test asks whether the output betrays the fact that there is a machine behind it. *Columba Loquens* asks the scientific equivalent: can the finished research artefact itself tell us whether the empirical world it reports was actually observed or synthetically constructed? The programme does not claim to have settled that question through a blinded experiment with human reviewers. It poses the test by constructing the artefact.

The implications reach beyond the pigeons. If an AI-generated dissertation can eventually no longer be distinguished from a conventionally produced dissertation by inspecting the dissertation itself, what exactly is a prohibition on AI authorship supposed to prohibit — and how could such a prohibition be enforced? Detection at the level of prose becomes a fragile foundation for regulation if the origin of that prose is no longer reliably visible in the product.

The problem then shifts upstream. Instead of trying to infer authorship from linguistic traces, scientific integrity has to pay greater attention to provenance, transparency and responsibility. Where did the data come from? Which parts of the process involved AI? Can consequential decisions be reconstructed? Who controlled the sources and the evidence? And who is responsible for the claims that survived the process? The distinction that governed the pentalogy becomes important again: textual production can be delegated without delegating scientific responsibility.

Here the unusual documentation of the programme matters. The five manuscripts are accompanied by extensive interaction histories, intermediate documents and methodological reflection reports that reconstruct how the work was produced, criticised and verified. Conventional publications normally show the finished artefact while most of the production process disappears behind it. Here that production history itself becomes inspectable material. It is possible to look back not only at what the final text says, but at how decisions developed, where output was rejected, which forms of control were introduced and where those controls themselves failed. The result is a degree of forensic transparency that ordinary publication practices rarely make possible.

All of this leads to a paradox that may be particularly important for universities. The easier it becomes to generate expert-looking academic work, the more important genuine expertise may become. A person who has never learned to construct an argument, weigh sources, recognise conceptual slippage or maintain a complex line of reasoning can now generate prose that looks as if those abilities were present. What may be missing is precisely the ability to tell when the result is wrong.

The **AI paradox** can therefore be stated more sharply: one may be able to use artificial intelligence fruitfully, critically and originally in scientific knowledge production only after first acquiring enough academic training without AI. It is a proposition emerging from these experiments, not a universal law. But it has obvious consequences. If students outsource the very work through which academic judgement is formed, they may learn to use AI before they have learned to judge its output. That does not make AI an argument for less education. It may be an argument for taking intellectual formation more seriously.

It also suggests why the opportunities of generative AI may so far have received less attention than its risks. Individual uses of AI in writing, summarising or searching are now commonplace. What appears to remain highly unusual — and may be unprecedented in this documented form — is the combination attempted here: several complete dissertation-scale studies across substantially different scientific practices, AI-generated dissertation prose under explicit human curatorship, internal AI review and verification, detailed process reconstruction and an overarching analysis of what the experiment produced. Any claim to being first should be treated cautiously. But the unfamiliarity itself matters. If few researchers have pushed generative AI this far through complete research processes, we should not be surprised that much of the academic debate still has a clearer view of the dangers than of the productive possibilities.

Those possibilities are substantial. The programme suggests three in particular. The first is obvious: **speed**. A considerable amount of human production labour can be removed while academically substantial work remains possible. The second is **synthesis across boundaries**. When the cost of processing and producing text falls dramatically, larger multidisciplinary constructions become easier to attempt. That may provide a counterweight to an academic culture in which knowledge is often divided into ever smaller, safely bounded contributions. Less human labour spent manufacturing text can free capacity for the larger questions, connections and syntheses for which researchers are ultimately valuable. The rapid production of the English versions points in the same direction: lower translation costs can also make it easier to bridge academic language communities.

The third opportunity is **verification**. AI does not merely make more production possible. Properly organised, it also makes far more extensive scrutiny possible. A second model can challenge a manuscript while it is being written. Source use can be examined systematically rather than sampled. Large populations of generated artefacts can be checked at a scale beyond realistic manual inspection. Verification becomes a production function in its own right.

This is the hopeful conclusion of the experiment. It can be done. It can, with appropriate controls and human responsibility, be organised in a way that proved workable in practice. And it does not imply the end of the academic researcher.

It implies a different division of labour.

The scientist need not spend the same proportion of time manufacturing prose, manually recombining information or performing checks that machines can repeat almost without cost. Human capacity can move towards choosing questions, designing research, constructing and challenging explanations, deciding which evidence matters and taking responsibility for the result. AI does not have to impoverish scientific work; used well, it may allow researchers to concentrate more of their time on the parts of scientific work where human judgement matters most.

That is also why the programme ends with a proposal rather than a prohibition: the **Socratic Oath**, modelled on the Hippocratic Oath. Its principle is straightforward. If AI acquires a larger share of scientific production, human responsibility must become more explicit, not less. The shortened version of the oath requires transparency about AI use, continuing human accountability, verification of sources and data, explicit attention to model limitations and biases, and scientific claims that remain open to scrutiny and testing.

The sequence matters. First comes the provocation: AI can already take over far more academic production than the conventional image of the researcher assumes. Then comes the opportunity: very large savings in human labour, faster synthesis, greater scope for crossing disciplinary boundaries and forms of continuous quality control that would otherwise be prohibitively expensive. Then comes the practical lesson: the limitations encountered were not negligible, but they proved sufficiently manageable to build working procedures around them. And finally comes the answer: exploit the new productive capacity, but design the system so that responsibility, provenance and judgement remain visible.

The debate about AI in universities therefore need not be framed as a choice between technological enthusiasm and defensive prohibition. There are uses that should be prohibited and uses that should be regulated. But there is a third option that deserves much more serious attention.

Exploit it. The books are there. So are the records of how they were made.

The question is no longer only what generative AI might do to science. We can begin examining what happened when it was actually allowed to become the writer — while the human researcher remained responsible for the science.

Note on the author and documentation. Dr Remco Schimmel received his PhD from the University of Twente in the Netherlands in 2007 and has since worked for many years as an independent researcher. The Dutch-language versions of the complete *AI-Driven Knowledge Production* pentalogy — the five dissertations together with the accompanying methodological reflection reports documenting their production processes — have already been registered with the Benelux Office for Intellectual Property (BOIP). Their English-language equivalents have likewise been deposited with BOIP and made publicly accessible through its i-DEPOT system**.** The reflection reports provide a separate methodological record of the human–AI production process, including its failures, corrections and verification procedures**.** Readers can therefore consult both the research outputs and the documentation of how they were produced. **For full documentation:** see https://research.cyber-busters.com/