# Autonomous biodiversity credits on the horizon?

Joseph Millard[1]*, Samira Barzin[2], Peter McCann[3], Lynn V. Dicks[1]

[1] Department of Zoology, University of Cambridge, Cambridge, UK

[2] Environmental Change Institute, University of Oxford, Oxford, UK

[3] UK Centre for Ecology and Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford, UK

*Corresponding author. Email: jwm57@cam.ac.uk

## Abstract

Biodiversity credits are being pushed as a means to fund nature conservation. Much of the debate around credits has concerned additionality, leakage, and permanence, and the extent to which biodiversity can be captured in an individual unit. As AI models continue to develop, however, technology could create a new kind of loss-of-control problem for biodiversity credits. In this Perspective, we express a concern that agentic AI could lead to the development of high-risk autonomous credits, where nation states cede control of nature restoration to the price of digital assets that are paid out with no human in the loop. We define an autonomous credit as a financial asset for nature quantified remotely through an agentic AI combined with a spatial finance model and real-time programmatic payment. Such autonomous credits do not yet exist, but appear to be in development. We highlight three technological step changes that signal autonomous credits are on the horizon, and then suggest three approaches that might help mitigate loss-of-control and maximise the benefits of credit technology. Given the development of agentic AI, where autonomous credits are connected to the critical infrastructure of food production and finance, these credits will need guardrails.

## Main

Biodiversity credits are being pushed as a means to fund nature conservation[1]. A certificate representing a positive change in biodiversity additional to what would have occurred otherwise, the appeal of biodiversity credits is that measurable and unambiguous outcomes for biodiversity can be bought and sold transparently on a market[2]. Such appeal means biodiversity credit methodologies have grown. As of September 2023 there were 34 different credit initiatives, all aiming to solve some dimension of undesirable biodiversity change. Of the single-country credits, six have been implemented in Oceania, four in Europe, three in South America, one in Africa, and none in Asia[1]. Most of these credits are issued for metrics combining some measure of species and ecosystems[1]. The Convention on Biological Diversity now includes these biodiversity credits as a global mechanism for urgent action on biodiversity loss [Global Biodiversity Framework Target 19; Convention on Biological Diversity 2022].

Credit development makes research concerning credits important. To effectively address biodiversity change, credits need to be provably valid with respect to additionality, leakage, and permanence[3], and to certificate meaningful outcome measures for biodiversity. For this latter point, attempts to derive universal units of nature for credits have proved controversial[2]: it is not possible to reduce biodiversity into one measure representative of its multi-dimensionality. By definition one dimension will not represent many dimensions. Meaningful universal units of nature is likely to remain an unsolvable

problem. However, given this diversity of opinion has probably always been true[4], lack of agreement might not stop the further development of credits.

As AI technology continues to develop, another problem could emerge. In this Perspective, we express a concern that autonomous credit markets, combining the agentic (i.e., able to act independently according to some purpose) retrieval of data, a remotely monitored spatial finance model, and programmable payment, could precipitate future loss-of-control scenarios (i.e., AI technologies that a human creates, and then finds itself unable to control). Specifically, we argue that if demand is driven up, for example through further legislation requiring the issuing of credits, the price of credits may rise. Costs on businesses given this legislation will create a new incentive for the development of autonomous systems for the monitoring, quantification, and issuing of credits. If these systems are created, autonomous credit markets—for both biodiversity and carbon—could mean nation states ceding control of nature restoration to the price of digital assets, issued and paid out with no human in the loop. Importantly, once an entity has created an autonomous credit monitored and paid out by an agentic AI, if deliberate safeguards are not implemented, it might prove difficult for that entity to maintain control of the software and hardware issuing that credit. Without stopping rules, loss-of-control might continue irrespective of any new legislation. High-risk autonomous credits such as this do not yet exist, but steps towards them appear to be occurring (see Table 1).

To emphasise, our argument is not against automation or some individual AI algorithm, it is not about biases or metrics, it is not a judgment on the merits of remote monitoring or field ecology, and it is not about the valuation of biodiversity, although many of these problems may be exacerbated by autonomous credit systems. We only warn that autonomous digital assets for nature—where an unambiguous link is made between remote monitoring, the structure of the land, and the price of a credit, agentically with no human involved—may be high-risk with respect to loss-of-control, even when these credits are optimising a measure of biodiversity we can agree is meaningful. This is a distinct and new technological problem.

In summary, first we provide a brief background on the development of agentic AI and AI safety and control, before describing a set of characteristics where loss-of-control in the economy might be more likely. Second, we describe that developments in AI and remote monitoring are developing rapidly, potentially enabling new forms of autonomous biodiversity credit. We argue that autonomous biodiversity credits will possess many of the characteristics of high risk regions of the economy for loss-of-control. We then highlight three step changes that signal autonomous credits are on the horizon, and assess a set of currently developing credit markets—for both biodiversity and carbon—against each of these three step changes. Third, we briefly explain why loss-of-control risks have not yet materialized for the more mature carbon credit markets. Fourth, we describe how to think about the risks potentially emerging from an autonomous biodiversity credit, and describe why these credits might be problematic. We then describe three characteristics inherent to an autonomous biodiversity credit that might reduce the likelihood of loss-of-control. We finish by suggesting three approaches (agentic AI safeguards, checkpoints involving a person in the physical realm, and regulation on outcomes we might not want) that might help both mitigate the risks of loss-of-control, and maximise the potential of autonomous credit technologies.

Mitigating in advance of developing technology is a core principle of AI safety and control[5]. However unlikely loss-of-control might seem, developers, governments, and regulators should be applying the precautionary principle and mitigating now in advance of developing technology. As far as we are aware there have been no papers on loss-of-control risks for AI in conservation (although see[6] for a science-fiction book on a similar idea). Our field is considering advanced AI for biodiversity credits[7], interacting with the critical infrastructure of both the financial system and food production. We are doing so in a space that has little regulation whilst the power of agentic AI continues to grow.

**Advanced AI will create unfamiliar risks**

For over 60 years researchers have recognized that a problematic AI might only need a goal, and the freedom to achieve that goal of its own accord. A classical thought experiment concerns a paperclip[8]. Suppose that an advanced AI is given the goal of creating paperclips, but no other specification as to how it should achieve that goal. This goal seems innocuous, but an advanced AI might reason that humanity and the resources it requires is a barrier to the fulfillment of this goal. As a result, diverting significant resource from the maintenance of human civilization becomes the most effective means of manufacturing paperclips. Further, since being switched off would prevent the advanced AI from fulfilling its goal, it might also subvert any human efforts to do so. Effectively, the assignment of an innocuous goal leads the AI to create subgoals for the fulfillment of that broader goal. Since the advanced AI has been given the freedom and capability to make its own decisions, the consequences of these subgoals can be difficult to anticipate. Of course, this example is deliberately provocative, but it demonstrates that perverse and uncontrollable outcomes can theoretically emerge from an advanced agentic AI where only seemingly harmless tasks are assigned. An advanced AI pursuing its own goals uncontrollably in this manner is described in the AI safety literature as "loss-of-control"[9].

Recent developments in agentic AI make the debate on loss-of-control no longer theoretical[9]. Since the release of ChatGPT in late 2022, AI technologies have moved quickly. Understanding why is conceptually simple. The capability of modern AI systems are broadly predictable on the basis of three parameters: computational power, data, and model size[10]. As AI companies (e.g., OpenAI, Anthropic) scale these parameters, model capabilities have grown[11]. In 2022, GPT3, the model underpinning the first iteration of ChatGPT, could respond to prompts in the manner of a human, providing reasonable responses and generating usable computer code[12]. Now, highly advanced AI systems exist that can carry out most of the role of an experienced software engineer[13] (e.g., Claude Code). AI systems exist that can initiate many separate agents (i.e., distinct AI instances that can act independently according to some purpose), and then autonomously follow goals across many domains that would ordinarily take a human many hours[14]. These systems can identify when they are under experimental observation during safety testing, and then use this information to mislead evaluators[15].

Developing technology means warnings about the risks posed by advanced AI systems are increasing[9]. An international panel of AI experts described loss-of-control risks as greatest where advanced AI systems are assigned goals interacting with critical global infrastructure, such as the financial sector[9]. The highest risk AI systems will be capable of eluding or undermining human control, will be able to use those capabilities to achieve high risk goals that are likely to result in loss-of-control, and will have been deployed by humans to enact those high risk goals[9]. Each of these three points sounds abstract. This abstraction is deliberate. Abstraction is required because loss-of-control has never been encountered as a problem before. Abstraction captures the deep uncertainty of the many contexts in which loss-of-control might happen.

Our own view is that loss-of-control risks resulting from advanced agentic AI may more likely emerge from unexpected regions of the economy (i.e., where the need for AI regulation is not immediately obvious), which intersect with multiple forms of critical infrastructure. Regions of the economy more likely to experience loss-of-control will have high economic incentive on establishing as an early mover, high amenability to full autonomy, the potential to propagate multiple forms of risk across spatial and temporal scales, and a philosophy on low human intervention. The workforce will probably be both unaware it is developing an agentic AI, and unaware of debates on AI safety and loss-of-control. Many of these points will be true of new forms of autonomous biodiversity credit technology.

**Nature finance technology is developing rapidly**

As the capability of general purpose AI has grown, technology for remotely sensing Earth's surface has also developed rapidly[16]. Measures of biodiversity and carbon can now in principle be monitored from space via satellites, from the air through drones, and from the ground via eDNA and acoustic monitoring, each connectable via an internet of things[16]. New geospatial intelligence legislation has been introduced[17], allowing the market to purchase satellite data, as a result opening up new opportunities for the connection of remote monitoring to broader autonomous systems of global finance and food production. Edge computing is developing, allowing advanced AI models to be run cheaply at the point at which data is sensed[18]. And more broadly for remote monitoring technology, new forms of self-supervised Earth-observation model, in which large quantities of satellite data are compressed into multi-dimensional strings of numbers called embeddings, are now able to detect many forms of variation at the Earth's surface[19].

Credits for biodiversity and carbon can benefit from agentic AI and remote monitoring. Agentic AI and remote monitoring can address two core problems for credits. First, the cost of registering projects is high[20], with documentation for the registration of credit projects often extensive[20]. Agentic AI may prove applicable in registering projects and completing documentation. Second, ecological monitoring in the field is labour intensive and expensive. Credit developers may reason that autonomous remote monitoring technologies can both bring down the cost of quantifying credits[21], and provide a traceable digital signature of biodiversity change.

Three relevant step-changes for credit technology are occurring now (Figure 1). If regulation pushes up the price of credits, economic incentive and these three technological step-changes may drive the development of autonomous systems for the registering, quantification, and issuing of credits. All three of these technological step-changes are happening independent of any legislation on credits. First, agentic AI LLMs (large language model) are now facilitating the autonomous retrieval of biodiversity data from the academic literature and biodiversity data APIs[22] (Application Programming Interface, i.e. a programmatic request to a database), allowing the autonomous and repeated improvement of conservation databases. The time horizon on task length for agentic AI is growing[14], meaning autonomous AI agents are increasingly able to perform longer tasks with no human involvement. Agentic AI systems are becoming increasingly powerful and difficult to evaluate[15]. Second, financial models built upon satellite data will allow the quantification of carbon and some measure of biodiversity from space[23]. Third, programmable payment will allow autonomous payment following the satisfaction of some pre-set criteria[24]. In combination, these three step-changes will mean credit developers could feasibly quantify, validate, and issue global biodiversity credits, theoretically in real-time, fully autonomously, and at low labour cost. The technology for all three of these step changes is close. A credit such as this is more speculative politically and economically than it is technologically. Agentic AI is the core technological unknown, with wide uncertainty on the capability of these systems on only a 2-3 year timeline[9].

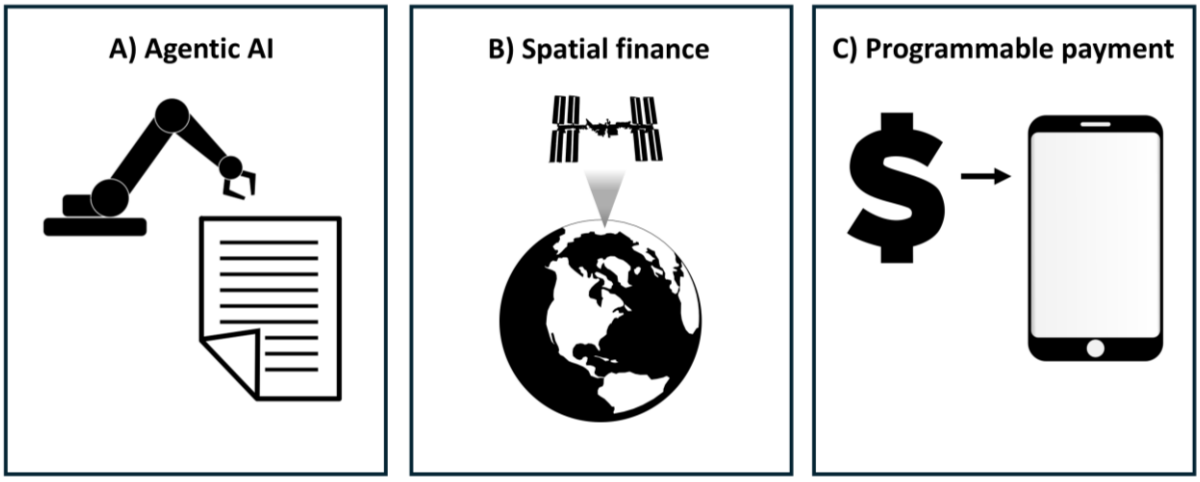


**Figure 1.** Three technological step changes on the horizon that relate to the development of new forms of autonomous biodiversity credit. A) Agentic AI (i.e., software that can autonomously and independently carry out tasks on a computer according to some goal); B) Spatial finance (i.e., a financial model in which satellite spatial data is used to quantify an asset remotely); and C) Programmable payment (i.e., autonomous payment following the satisfaction of pre-set criteria). Credits combining an agentic AI with a spatial finance model and programmable payment may be high risk with respect to loss-of-control, given the way in which it would hand conservation decisions to a single joined-up autonomous system.

Even with little demand for credits, there are at least two biodiversity credits that appear to satisfy two of our step changes (Table 1). MERIT and GainForest both appear to connect a remote monitoring system to a programmable payment. No forms of digital asset for biodiversity or carbon yet include the agentic collation of data, although we know that this is the direction of travel at least for evidence synthesis in conservation science[22]. With respect to our three step changes, carbon credit technology now appears to be lagging behind biodiversity credits (Table 1). As far as we are aware there are no carbon credits combining a spatial finance remote monitoring model with a programmable payment. Possibly this results from criticism of carbon credits[25], making credit developers wary.

**Table 1.** Autonomous biodiversity and carbon credits in development, measured against three criteria: Agentic AI, spatial finance remote monitoring (i.e., a financial model in which satellite spatial data is used to quantify an asset remotely), and programmable payment. All credits here have at least some partial development towards either spatial finance remote monitoring or programmable payment, identified through an exhaustive search using Google Search and ChatGPT, carried out in October 2025. 'x' indicates any criterion that appears close to being implemented in at least some form for that credit.

| **Biodiversity credit** | **Agentic AI data retrieval** | **Spatial finance remote monitoring** | **Programmable payment** |
|---|---|---|---|
| MERIT | | **x** | **x** |
| GainForest | | **x** | **x** |
| RegenNetwork | | **x** | |
| Veritree | | **x** | |
| Treegens | | | **x** |
| Treejer | | | |
| Silvi | | | |
| Epoch Blue | | **x** | |
| **Carbon** | **Agentic AI data retrieval** | **Spatial finance remote monitoring** | **Programmable payment** |
| Fedrok | | | **x** |
| Token ovate | | | |
| Northern Trust Carbon Ecosystem | | | |
| CarbonCoins | | | **x** |
| Climera | | | |
| KlimaDAO | | | |

**So what's the problem?**

Biodiversity credits combining an agentic AI with a spatial finance model and programmable payment could mean conservation decisions handed to a single joined-up autonomous system. Without due care, a system such as this where a link is made between remote monitoring, the quantity of restoration, and the price of a credit, all with no human in the loop, could be high risk with respect to loss-of-control. Prior precedent on cryptocurrencies helps reveal why.

Cryptocurrencies are digital currencies reliant on no central authority[26]. These digital currencies are not generated by a monetary authority or bank, but rather "mined" through the execution of a program across a distributed network of computers[26]. Vast networks of computers now exist that exclusively mine cryptocurrency[27]. Cryptocurrencies have demonstrated two things relevant for an autonomous biodiversity credit: cryptocurrencies have shown that unregulated digital assets can be generated programmatically across a distributed network (i.e., by the execution of a programme), and that these assets tend towards high volatility, with rapid changes in price[28].

Suppose now that some autonomous biodiversity credit is designed and released for the restoration of forest. It doesn't necessarily have to be a cryptocurrency, but it is "mined" by a landowner when a remote monitoring system confirms uplift of natural forest in their area of land (i.e., uplift of natural forest causes the execution of a programme which autonomously pays that landowner a credit), and it is built end-to-end into an agentic AI system. As we have described, the technology for this form of credit is close. An autonomous biodiversity credit could inherit both distributed programmatic generation and the volatility of digital currencies, but differ in an additional important respect. Unlike a conventional digital asset, an autonomous biodiversity credit does not exist in digital space alone, but is also linked directly to the structure of the land via the state of biodiversity. As a result, autonomous credits could propagate financial risks typical of cryptocurrencies, but also new forms of risk linked to the use of land. Deliberate price inflation of an autonomous credit, for example, could feasibly cause nature restoration projects to displace farmland, leading to increases in food price. The difference between an autonomous credit and a conventional credit, is that once price inflation has been recognised as a problem, the autonomous credit software might already exist across a distributed network of computers. In other words, the issue for credits such as this is that they risk creating problems we have seen previously for credits, but in such a way that fixing those problems then becomes not possible. A worst case scenario, for example, could be an agentic autonomous credit requesting publicly available satellite data. Preventing the issuing of that credit might mean removing all online public access to that data stream, such that there is no way for an agentic AI to access it, and equally no way for any human to access it either.

An autonomous biodiversity credit would inherently possess three characteristics, which might at least reduce the likelihood of a worst-case loss-of-control scenario: first, nature restoration projects are relatively slow given the physical constraints of plant growth, which may provide a natural inertia such that price volatility is low; second, downstream effects in the market might help to equilibrate the price of credits relative to profits generated from agriculture; and third, the purchase and sale of land is slow, with legal frameworks for the exchange of land differing significantly between states. Although all of these points are true, they provide no verifiable guarantee (see[5] for the importance of verifiable proof with respect to AI safety and control)—all three only reduce the likelihood of worst-case scenarios. None provide a guarantee for the behaviour of an agentic AI. Further, for the third of these points, owners of land might see greater profit in switching quickly from food production to nature restoration, meaning exchange of land might not be needed for the precipitation of worst-case scenarios.

However unlikely the development of autonomous biodiversity credits might seem, where the stakes on loss-of-control are potentially high, it makes sense to apply the precautionary principle and mitigate in advance of developing technology[5]. This is a general principle of AI alignment and control[5]. There are probably three core ways in which we can maximise the benefits of autonomous biodiversity credits and

minimize the risks. The first way is to ensure that agentic AI built into autonomous biodiversity credits has safeguards such that any credit can be demonetised if needed (i.e., the developer maintains control over the software and hardware infrastructure after it has been released). This could involve the development of check-and-balance AI systems[29], monitoring the chain-of-thought of the agentic AI (i.e., the thought process of the model, see[30]), or deliberately building systems such that they are highly responsive to human input[9,31]. The second way is to implement safety checkpoints in the physical realm. Autonomous AI credits built across multiple steps, from the quantification of a credit all the way to issuance and payment, will be high risk. In other words, one way to mitigate risk is to have a person involved in at least some stage of the process of the registration or selling of a digital asset for nature. Having a person involved will help to dampen credit volatility, and means there is always somewhere legislation can intervene. The third way, which is not necessarily mutually exclusive, is to regulate on goals we might not want, which can be anchored to measurable outcomes for the biosphere. If there is regulation on goals and these are built into an agentic AI, we can at least know that an autonomous credit will not create perverse outcomes (e.g., large homogeneous plantations).

**Conclusion**

Forms of autonomous biodiversity and carbon credit could have a role to play in financing nature restoration and conservation. Technology can free credits from the labour costs of monitoring, and create transparent mechanisms for the certification and exchange of these assets. But agentic AI technologies applied to credits could also propagate new forms of loss-of-control risk from themselves. Here we highlighted three technological developments that signal autonomous digital biodiversity credits are on the horizon (agentic AI, a remotely monitored spatial finance model, and programmable payment). We then suggested three ways in which loss-of-control risks from autonomous credits might be mitigated (agentic AI safeguards, checkpoints involving a person in the physical realm, and regulation on outcomes we might not want). There may be a path to autonomous credits that are provably valid with respect to additionality, leakage, permanence, and controllability, but developers and regulators should proceed with caution and mitigate in advance of developing technology.

## References


1. Wunder, S. *et al.* Biodiversity credits: learning lessons from other approaches to incentivize conservation. Preprint at https://doi.org/10.31219/osf.io/qgwfc_v1 (2024).
2. Wauchope, H. S. *et al.* What is a unit of nature? Measurement challenges in the emerging biodiversity credit market. *Proc. R. Soc. B Biol. Sci.* **291**, 20242353 (2024).
3. zu Ermgassen, S. O. S. E. *et al.* Five rules for scientifically credible nature markets. *Nat. Ecol. Evol.* **10**, 181–192 (2026).
4. Sandbrook, C., Scales, I. R., Vira, B. & Adams, W. M. Value Plurality among Conservation Professionals. *Conserv. Biol.* **25**, 285–294 (2011).
5. Russell, S. *Human Compatible: AI and the Problem of Control*. (Penguin UK, 2019).
6. Beauman, N. *Venomous Lumpsucker*. (Soho Press, 2022).
7. Ford, H. V. *et al.* A technological biodiversity monitoring toolkit for biocredits. *J. Appl. Ecol.* **61**, 2007–2019 (2024).
8. Wiener, N. Some Moral and Technical Consequences of Automation. *Science* **131**, 1355–1358 (1960).
9. Bengio, Y. *et al.* Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845 (2024).
10. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
11. Achiam, J. *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
12. Brown, T. *et al.* Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
13. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7.
14. Lovely, G. AI could soon tackle projects that take humans weeks. *Nature* https://doi.org/10.1038/d41586-025-00831-8 (2025) doi:10.1038/d41586-025-00831-8.
15. Lynch, A. *et al.* Agentic Misalignment: How LLMs Could Be Insider Threats. Preprint at https://doi.org/10.48550/arXiv.2510.05179 (2025).
16. Besson, M. Towards the fully automated monitoring of ecological communities. *Ecol. Lett.* **25**, 2753–2775 (2022).
17. *Geospatial Data Act Online Version — Federal Geographic Data Committee*. (2018).
18. Vuilliomenet, A., Jones, K. E. & Wilson, D. Future of Edge AI in biodiversity monitoring. Preprint at https://doi.org/10.48550/arXiv.2602.13496 (2026).
19. Feng, Z. *et al.* TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis. Preprint at https://doi.org/10.48550/arXiv.2506.20380 (2026).
20. Croci, E., Lucchitta, B. & Cusa, M. Biodiversity credits schemes: a comparative analysis. *J. Clean. Prod.* **523**, 146382 (2025).
21. Schöttker, O. *et al.* Monitoring costs of result-based payments for biodiversity conservation: Will UAV-assisted remote sensing be the game-changer? *J. Nat. Conserv.* **76**, 126494 (2023).
22. Reynolds, S. A. *et al.* The potential for AI to revolutionize conservation: a horizon scan. *Trends Ecol. Evol.* **40**, 191–207 (2025).
23. Caldecott, B. Spatial finance: practical and theoretical contributions to financial analysis. *J. Sustain. Finance Invest.* (2022).
24. Weber, I. & Staples, M. Programmable money: next-generation blockchain-based conditional payments. *Digit. Finance* **4**, 109–125 (2022).
25. Macintosh, A. *et al.* Carbon credits are failing to help with climate change — here's why. *Nature* **646**, 543–546 (2025).
26. Milutinovic, M. Cryptocurrency. *Ekon. J. Econ. Theory Pract. Soc. Issues* **64**, (2018).

27. Wendling, M. The Bitcoin hum that is unsettling Trump's MAGA heartlands. *BBC News* https://www.bbc.co.uk/news/articles/cx2qg6e03l2o (2025).
28. Sözen, C. Volatility dynamics of cryptocurrencies: a comparative analysis using GARCH-family models. *Future Bus. J.* **11**, (2025).
29. Xiang, Z. *et al.* GuardAgent: Safeguard LLM Agents by a Guard Agent via Knowledge-Enabled Reasoning. Preprint at https://doi.org/10.48550/arXiv.2406.09187 (2025).
30. Korbak, T. *et al.* Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. Preprint at https://doi.org/10.48550/arXiv.2507.11473 (2025).
31. Potham, R. & Harms, M. Corrigibility as a Singular Target: A Vision for Inherently Reliable Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2506.03056 (2025).

**Acknowledgments**

Thanks to Tom Swinfield, Chris Sandbrook, Tom Frederick Johnson, and Mike Harfoot for comments and discussions on earlier drafts. Thanks also to three anonymous reviewers who provided useful comments on a first round of submission. JM is funded by the Leverhulme Trust and the Isaac Newton Trust on an Early Career Fellowship.