# DESTINATION UNKNOWN: AI AND IP IN THE DIGITAL ECONOMY

Generative AI has revealed a blind spot in liberal private law, of which intellectual property law is illustrative. Copyright's statutory verbs fail to describe what models actually *do* with training data, while contract and property doctrines have enabled enclosure. The result is a political economy of double extraction: Web 2.0's attention harvest and model developers' appropriation of human traces to generate synthetic substitutes. The rise of "AI slop" makes this approach unsustainable, as authentic human contributions become scarce and newly valuable. This paper argues that liberal private law can be refitted for the digital economy, but the challenge is no longer doctrinal fine-tuning.

Jason Grant **ALLEN**[1]
*BA & LLB (Hons) (UTAS), LLM (Augsburg), GDLP (College of Law Australia), LLM (UTAS), PhD (Cambridge);*
*Associate Professor of Law and Lee Kong Chian Fellow, Yong Pung How School of Law, Singapore Management University;*
*Director, SMU Centre for Digital Law;*
*Adjunct Associate Professor, School of Law, University of Tasmania.*

## I. Introduction

1 Generative artificial intelligence ("GenAI") models are trained on vast *corpora* of text, images, audio, and video; they "learn" patterns and produce outputs that astonish in fluency and fidelity. Beneath this simple description, however, lies a set of legal puzzles, including some live issues in the law of intellectual property ("IP"). This paper focuses on one subset of those issues, namely those relating to the possibility of *input-side infringement* by using copyright protected material in training datasets. This set of issues has (a) a *doctrinal or legal-technical aspect*, at the core of which is a basic question: Is the use of copyright protected materials to train a model a copyright infringement? The answer to this

1 Thanks to Jake Goldenfein and Huijuan Peng for their comments on the draft. Thanks also to the various participants of the Singapore Management University colloquium on "Current Issues on AI & Intellectual Property in the Asia-Pacific" ("Colloquium") held over 2024 and 2025 with support from Google. The discussions in that Colloquium are the inspiration for this analysis. As usual, all errors remain the author's own.

question is binary (yes or no), but not always easy. It differs according to the law of the relevant jurisdiction, and currently, there is a good deal of litigation pending on it, particularly in the US. There is also, however, (b) a *normative or legal-policy aspect*. Repeatedly, over the course of our Colloquium on "Current Issues on AI & Intellectual Property in the Asia-Pacific[2] ("Colloquium") from which this paper draws its inspiration, discussions on points of IP law bled into discussions about the bigger picture – even though attempts were made to insulate the "big questions" to a dedicated session. The legal-policy question is whether the law as it stands does what we expect it to do – and, if not, how it might need to change.

2 The two aspects of the input-side infringement puzzle are related, and the doctrinal aspect provides the necessary set-up to properly frame the policy aspect. The author's emphasis, however, is on the latter. This paper proceeds with copyright's semantic ambiguities before widening to the broader failure of private law to grapple with data. The key question is: How should we understand what these systems *do* with training data? Copyright law's vocabulary does not necessarily map directly onto the parametric operations of large language models ("LLMs"). Further, private law more broadly has failed to provide categories that grasp "data" *per se* as a legal object. This failure is not trivial. Liberal private law, in both its doctrinal architecture and its normative ideals, aims to secure autonomy and dignity through consensual transactions, primarily using the building blocks of property and contract. If data is central to the economy of the digital age, the inability to categorise or allocate it properly undermines its promise. It is useful to contrast the liberal private law approach with the explicit recognition of data as a "factor of production" in the People's Republic of China ("PRC"). This is a framing somewhat alien to liberal sensibilities, but it is illuminating in its clarity about data's role in the *political economy* emerging around AI. At the least, taking this alternative approach seriously informs a serious reformist agenda for those of us working within the liberal tradition. From there, the argument tracks the political economy of digital platforms: Web 2.0's enclosure and "enshittification"[3] and GenAI's double extraction of value. Against this backdrop, this article explores the rise of "AI slop" and the looming scarcity of authentic human contributions, framed normatively through John Rawls's "Original Position" thought experiment and

2 Hosted by the Centre for Digital Law at the Singapore Management University and supported by Google, the Colloquium was convened by Associate Professors Jason Grant Allen and Saw Cheng Lim over three sessions spanning the years 2024 and 2025.

3 Merriam-Webster, "Enshittification: What does enshittification mean?", *Merriam-Webster.com* <https://www.merriam-webster.com/slang/enshittification> (accessed 14 October 2025). See also para 34 below.

institutionally through Katharina Pistor's framing of the law's role in constituting the furniture of the economic world.[4] A reformist conclusion is reached: liberal private law must change if it is to meet the challenges of the digital economy.

## II. Setting the scene

3 At a recent panel on AI and IP, a law professor asked a practitioner a simple question: *If we had known in advance how things would unfold, would we have designed a better licensing regime for copyrighted works?* This will be referred to as the "Professor's Question". The "Practitioner's Response" followed a familiar pattern: Individual works, in isolation, have negligible value – it is only in the aggregation of vast *corpora* that they acquire significance; the use of copyrighted material by an LLM is no different from that of a scholar who reads ten books and then writes one of her own, or a poet who borrows, parodies, and alludes to predecessors; above all, society gains from a permissive rather than restrictive approach to training – the greater the commons of data, the greater the spur to innovation. If we were to impose, for example, an overly restrictive interpretation of "fair use" or tie model developers up in intricate licensing webs, we would all lose out on this valuable innovation.

4 There is substance to this argument. But it also deserves to be interrogated. (The problems with the "regulation stifles innovation by private industry" rhetoric are only noted in passing here; the whole AI value chain, from chips to power plants, is a direct result of state – including, especially, defence – spending and industrial policy over decades. The idea that states need to just let the market do its thing seems disingenuous in a world where the daily news underscores the importance of AI in geopolitical competition.[5]) But this article is more concerned with the legal propositions marshalled in the Practitioner's Response. At the time of writing, news broke of Anthropic's landmark settlement with copyright owners – a potent vindication of the intuition behind the Professor's Question, as the company agreed to pay US$3,000

4 The "furniture" metaphor is not taken from Pistor but from Uskali Mäki: "Some nonreasons for nonrealism about economics". See *Fact and Fiction in Economics: Models, Realism, and Social Construction* (Uskali Mäki ed) (Cambridge University Press, 2002) at p 95. See also John Rawls, *A Theory of Justice* (Belknap Press, 1971); Katharina Pistor, "A Legal Theory of Finance" (2013) 41 *Journal of Comparative Economics* 315.

5 See further Gilad Abiri, "Mutually Assured Deregulation" (2025) *Stanford Technology Law Review* (forthcoming), <https://ssrn.com/abstract=5394963> (accessed 4 September 2025).

per book for 500,000 pirated books used in training data.[6] But this only covers books that were taken from pirate websites, and does not provide a complete answer to complex questions surrounding the US doctrine of fair use. There are other cases pending against other entities in other jurisdictions, so it is important to give the Practitioner's Response careful, if critical, consideration.

5 Take, for instance, the apparently semantic question of what verb best describes what LLMs *do* with training data. At one of the Colloquium roundtables, participants spent nearly an hour debating this, and not in vain. Some favoured technical verbs such as *train*, *encode*, *optimise*, *predict*. Others preferred anthropomorphic verbs like *learn*, *memorise*, *generalise*. Still others emphasised the data-handling aspect: *ingest*, *process*, *extract*, *represent*. And then there were verbs pointing to outcomes: *adapt*, *generate*, *model*. At first glance, this might look like semantic hair-splitting. But the stakes are real; as the Anthropic settlement shows, there are significant distributional questions here, and the available vocabulary does not always allow those questions to be addressed as directly as they should be.[7]

6 Copyright law works with a small cluster of defined verbs. For example, Anglo-American and Commonwealth copyright law focuses on verbs like *copy*, *reproduce*, *distribute*, *communicate*, *adapt*. Whether or not any given use of training data falls within copyright law's statutory perimeter depends on whether what a model does can be mapped onto these terms of art. Predictably, those in favour of permissive regimes choose verbs that diverge from copyright's vocabulary; those who want to find input-side infringement emphasise the overlap. Too often, these two stakeholder groups talk at cross-purposes. Although a consensus was not always reached during the Colloquium meetings, participants identified the problem-space and isolated the points in issue instead of talking past each other. It is legitimate for different stakeholders – for example, a model developer, a rights-holder organisation, and an academic – to take opposing positions on this question. That is how the legal system works – in particular, how the law is dragged into the present-day and

6 See Melissa Korn & Jeffrey Trachtenberg, "Anthropic Agrees to Pay at Least $1.5 Billion in Landmark Copyright Settlement", *The Wall Street Journal* (5 September 2025) <https://www.wsj.com/tech/ai/anthropic-to-pay-at-least-1-5-billion-inlandmark-copyright-settlement-with-authors-bfcdd57b> (accessed 12 September 2025).

7 See also Jake Goldenfein, "Privacy's Loose Grip on Facial Recognition: Law and the Operational Image" in *The Cambridge Handbook of Facial Recognition in the Modern State* (Rita Matulionyte & Monika Zalnieriute eds) (Cambridge University Press, 2024) ch 5.

applied to fact scenarios driven by different technologies to the ones that characterised the era in which it emerged.[8]

7 The most controversial point in this debate is whether the content of specific artefacts of training data – a book like *Harry Potter and the Philosopher's Stone*, say – is in *some* sense reproduced (for want of a better word) in the parametric space of the model. The fact that models can, when prompted in the right way, produce *verbatim* or near-*verbatim* passages suggests that the "shape" of the underlying work is inscribed somehow, somewhere in the weight-space of the network. It is true that if an individual asks ChatGPT today for the first paragraph of *Harry Potter*, it will likely refuse. But this is a matter of design choice – such that the bot "will not", not that it "cannot". The impossibility is an artefact of a safety layer, not a feature of the model's inner workings. From the standpoint of *input*-side copyright infringement, this matters: preventing *output*-side infringements tells us little about whether input-side acts fall within the scope of those specific, prohibited verbs.

8 This question is very much alive at the time of writing. In the weeks before publication, two conflicting cases were delivered that can be used to illustrate its currency, albeit without any fulsome analysis. *Getty Images v Stability AI Ltd*[9] concerned the well-known case of images bearing the "gettyimages" watermark. The model developer admitted that it had used Getty's images to train the model; for various reasons, the case ultimately turned on its defence going to the technical requirements of the Copyright, Designs and Patents Act 1988[10] s 22, which prohibits (under such circumstances as this) the importation into the UK of an "infringing copy" of a copyright work. Joanna Smith J held that the model was not an infringing "copy" of the copyrighted work because it did not "reproduce" that work – effectively, accepting the argument that even though the model was trained on copyright protected works, the model did not "reproduce" them. On this operative question, the judge held:[11]

> [I]t seems to me to be clear that an infringing copy must be a copy, as Stability submits; the essence of the infringement is that there has been an infringement of copyright by the reproduction of the work (including by its storage in any medium by electronic means) in any material form … [T]he dispute between the parties as it finally emerged in closing, really turns on whether an article

8 See, *eg*, discussion in Ben DePoorter, "Technology and Uncertainty: The Shaping Effect on Copyright Law" (2009) 157 *University of Pennsylvania Law Review* 1831; and Joseph Loewenstein, *The Author's Due: Printing and the Prehistory of Copyright* (University of Chicago Press, 2002).

9 [2025] EWHC 2836 (Ch).

10 (c 48) (UK).

11 *Getty Images v Stability AI Ltd* [2025] EWHC 2836 (Ch) at [597]–[600].

> whose making involves the use of infringing copies, but which never contains or stores those copies, is itself an infringing copy such that its making in the UK would have constituted an infringement. Taking the specific facts with which I am concerned, is an AI model which derives or results from a training process involving the exposure of model weights to infringing copies itself an infringing copy? ... In my judgment, it is not.

In the Munich Regional Court,[12] the German collective rights agency ("GEMA") sued OpenAI on the basis that the lyrics of certain songs had been memorised in the model parameters such that they could be reproduced *verbatim*. The 42nd Civil Chamber largely upheld GEMA's claims for injunctive relief, disclosure and damages, accepting the argument that copyrighted works could be embedded or embodied in a model's parameters in a way that infringed the German Copyright Act[13] ss 15, 16 and 19(a). In effect, the court held that both the memorisation in the language models and the reproduction of the song lyrics in the chatbot's outputs constitute infringements of copyright exploitation rights. OpenAI's actions were not covered by any exceptions (in particular covering text and data mining).[14]

9 This is why the choice of verbs is more than mere semantics. If copyright's vocabulary cannot be made to fit what models do, then there are two possibilities. One is that the law remains as it is, leaving model training outside its scope. The other is that the law changes – adopting new verbs for new concepts to capture technology-enabled acts that we think (as a matter of policy) need to be captured. Either way, the doctrinal debate quickly bursts its bounds and turns into a policy question. Put shortly: if copyright is not the right tool, what kind of legal framework would be? That is the point of departure for what follows. Copyright can be stretched or reinterpreted, but it was not designed for our data-driven economy. Along with other sub-regimes such as privacy and data protection law, IP is being used in the attempt to rein in "big tech" and regulate AI. To understand the real fault lines, there is a need to step back from copyright and look more broadly at private law in liberal

12 *GEMA v OpenAI* Case No 42 O 14139/24. See the unofficial translation and press release (11 November 2025) <https://ifrro.org/resources/documents/General/German_Court_OpenAI_Memory_Output_Infringe_Copyright_NOV25.pdf> (accessed 24 November 2025).

13 Copyright Act of 9 September 1965 (Federal Law Gazette I, p 1273) (Germany).

14 Although the author has not had a chance to read it, see also the recent argument by Rita Matulionyte that copyright law should recognise the "ingestion" of copyright protected works, if necessary reconceptualising "reproduction", in order to be fit for purpose in our century: Rita Matulionyte, "Reconceptualising the Reproduction Right in the Age of AI" (2025) *IIC: International Review of Intellectual Property and Competition Law* <https://doi.org/10.1007/s40319-025-01653-x> (accessed 24 November 2025).

legal systems, which is supposed to provide a *prima facie* neutral and fair game-space in which players secure autonomy and dignity through consensual transactions.

## III. Grasping at smoke: private law and the data-driven economy

10 There is a deeper problem here: Data itself does not fit any of private law's categories, and this misfit has profound consequences. Copyright's semantic ambiguities thus expose a more general problem. The difficulty is not simply that copyright's statutory verbs fail to map onto model training. The deeper difficulty is that *no branch of private law in liberal legal systems really grasps "data" at all*. And this is striking, because private law is supposed to be the domain that best expresses the liberal commitment to autonomy and dignity, and to construct an open-ended transaction-space in which things of value can be traded.

11 In a liberal order, contract and property do the essential work of enabling individuals and firms to pursue their ends through consensual transactions.[15] The law provides "building blocks" that are almost infinitely composable. Property secures entitlements; contract allows their transfer and exchange. Both institutions rest on the idea that individuals are rights-bearing agents who can shape their lives and projects by entering into agreements on equal footing. This is the normative ideal: private law as the infrastructure of autonomy, dignity, and – at least in its baseline form – equality. As Hanoch Dagan and Irit Samet explain:[16]

> … Property on this view is an autonomy-enhancing institution; one of the major legal tools that serve the primary commitment of every liberal polity to secure and facilitate people's foundational right to self-determination … As such, liberal property requires law to facilitate in each important area of human action and interaction a diverse set of stable frameworks of private authority … so that people can set up – on their own or with the cooperation of others – long-term plans.

12 But when this scheme is tested against the realities of today's data-driven economy, it begins to falter. Each sub-regime addresses a *part* of data's role in the data-driven economy, but it does not provide

15 See also Sebastian Benthall & Jake Goldenfein, "Artificial Intelligence and the Purpose of Social Systems" (2021) *AEIS 21: Proceedings of the 2021 AAAI/ACM Conference of AI, Ethics and Society* 3 <https://doi.org/10.1145/3461702.3462526> (accessed 8 October 2025).

16 Hanoch Dagan & Irit Samet, "The Beneficiary's Ownership Rights in the Trust Res in a Liberal Property Regime" (2023) 86(3) *Modern Law Review* 701 at 702, summarising Hanoch Dagan, *A Liberal Theory of Property* (Cambridge University Press, 2021).

a complete answer. That would not be so problematic if the various sub-regimes taken together made a coherent meta-regime, but that is not the case. Mindful that the author is merely passing as a "tourist" through many fields of specialism, a brief review is enough to make the point.

13 At first glance, *IP* looks like the natural home for data. Copyright, patents, database rights, and related regimes all concern intangible assets. But data sits awkwardly here. Much of the data used to train AI – personal data, transactional logs, scraped web pages – does not qualify as an "original work" or inventive subject matter. Database rights offer some protection in some jurisdictions, but not in most common law jurisdictions, and even there they protect the investment in *compiling* data, not the data points themselves. This makes IP a poor proxy for control over data in its raw or ambient forms.

14 *Property*, in liberal legal theory, secures exclusionary control over things. Yet data has never been recognised as a "thing" in the relevant sense. Anglo-American law has long been sceptical of extending property beyond land, chattels, and a delimited set of intangibles. Statutory interventions are possible – one could legislate that data is property – but common law courts and legislatures have generally resisted such moves. As Hu Ying writes, we find ourselves in the "rather odd situation" where data has huge commercial value, and we tend to think intuitively of "our" data belonging to us, but "we do not legally own" the data collected from and about us, and it is "not even considered 'property' in the first place".[17] In the Civilian context, the codified systems also struggle to take cognisance of data *per se* directly.[18] The result is that individuals cannot rely on property doctrine to secure control over the data they generate, either to prevent its use or to license it for any kind of rent.[19]

15 A common line of argument accepts the *prima facie* inappropriateness of data *per se* as an object of property rights and then posits the *quasi-proprietary* protections granted to certain types of data by privacy and data protection laws.[20] These relatively recent regimes

17 Hu Ying, "Private and Common Property Rights in Personal Data" (2021) 33 SAcLJ 173 at 173.

18 See Andreas Boerding *et al*, "Data Ownership – A Property Rights Approach From a European Perspective" (2018) 11(2) *Journal of Civil Law Studies* 324.

19 See also the discussion in Kean Birch, "Assetization As a Mode of Techno-Economic Governance: Knowledge, Education and Personal Data in the UN's System of National Accounts" (2023) 53(1) *Economy and Society* 15.

20 For a textbook example, see Law Reform Committee, Singapore Academy of Law, *Rethinking Database Rights and Data Ownership in an AI World* (July 2020) ch 3 <https://sal.org.sg/wp-content/uploads/2025/02/2020-Rethinking-Database-Rights-and-Data-Ownership-in-an-AI-World_ebook_0_1.pdf> (accessed 4 September 2025).

address some aspects of informational control, but they do not operate as private-law entitlements tradable in the market. They are regulatory overlays, framed as statutory compliance obligations or human rights. Consent, where it appears, is thin: checking a box on a form is no substitute for the thick consent envisioned in liberal contract theory. Nor do these regimes capture the productive dimension of data – the way aggregated data fuels AI and the wider digital economy.

16 This brings us to *contract*. Contract is the workhorse of the digital economy, but its operation here is paradoxical. On one side, the liberal ideal of contract – autonomous parties negotiating on equal footing – has collapsed into practices of adhesion: "clickwrap" agreements and terms of use that can be changed unilaterally. Users "agree" to sprawling platform terms they cannot be expected to read and which are non-negotiable, and which entrench rather than mitigate asymmetry. Consent in this context is largely fictional, and the substantive outcome is dictated by corporate power, not individual choice.[21] *Consumer law's* growing role in policing unfair terms is evidence of this failure: it shows that private ordering has not delivered autonomy and dignity, requiring regulatory correction at the margins. On the other side, it is through contract that most real-world data exchanges take place. Firms routinely "quasi-propertise" data by stipulating rights of access, use, exclusivity, and confidentiality.[22] Contract gives operational reality to data as an asset, even though property law does not formally recognise it as such. Licensing agreements, data-sharing arrangements, and platform application programming interfaces ("APIs") all embody this logic. In practice, contract both reifies data (*ie*, turns it into something "thing-like") and commodifies it – constructing *de facto* entitlements where formal law offers none.

17 This duality is instructive. Contract demonstrates how liberal private law simultaneously fails at its *normative project* (realising autonomy and dignity for individuals) and succeeds at one of its core *functional projects* (facilitating commodification and exchange of things of value). Predictably, the benefit of data commodification tends to follow asymmetries of economic power between users and platforms: Data is structured as a tradable asset for firms, while individuals remain locked into adhesion.

---

21 For further analysis in the privacy context, see, *eg*, Thomas D Haley, "Illusory Privacy" (2022) 98(1) *Indiana Law Journal* 75.

22 For an analysis of this dynamic focused on the human body, rather than data, see TT Arvind & Aisling McMahon, "Commodification, Control, and the Contractualisation of the Human Body" in *The Limits of the Market* (Elodie Bertrand, Marie-Xavière Catto & Alicia Mornington eds) (Mare & Martin, 2020).

18 Although it is perhaps adjacent to "private law", *competition law* is also a useful point of reference. Competition law does recognise data, but only indirectly, as a source of market power. Cases about search engines, app stores, or digital advertising occasionally treat control of data as an anti-competitive asset. But this is about market structure, not about the individual's relation to "their" data. It addresses downstream concentration without providing individuals with meaningful rights upstream. And as Alptekin Koksal argues, existing competition law tools are not currently effective in addressing market power in the data-driven economy.[23] The European Court of Justice decision in *Meta v Bundeskartellamt*,[24] for example, confirmed that antitrust enforcers may treat data protection breaches as a "vital clue" to abuse of dominance, but that link still operates through market-structure and conduct analysis; it does not create user-level data claims.[25] In other words, it regulates data's effects, but does not attempt to grasp "data" itself directly.

19 Taken together, these gaps leave liberal private law unable to deliver on its own promises. Individuals cannot treat data as "theirs", trade it on equal terms through contract, or rely on IP to secure control. Privacy laws offer some protection but not private ordering; competition law addresses structural concentration but not personal autonomy. The net effect is that data flows are governed less by liberal private law than by corporate practice, technical data architectures, and regulatory patches. The liberal vision of individuals shaping their lives through consensual transactions in a framework of autonomy and dignity does not extend to data. This is a major problem in an economy that is becoming more "data-driven" by the minute – particularly when seen in light of present concerns around labour market dynamics, workforce resilience, and disruption of the cycle of production and distribution (*ie*, humans working to produce outputs in exchange for wages, which they use to buy other outputs). This failure has implications at the personal level for "data subjects", and at the societal and economic level writ large.

### *A. Labour, data, and the personal dimension*

20 The failure of liberal private law to grasp data is not only a doctrinal gap. It also obscures the way data collection and AI training now capture forms of labour and expertise that were previously shielded from commodification. Knowledge and skill, once embodied

23 Alptekin Koksal, *Big Data and Competition Law: Market Power Assessment in the Data-Driven Economy* (Routledge, 2024).

24 *Meta Platforms Inc v Bundeskarellamt* Case C-252/21, EU:2023:537.

25 See Peter van de Waerdt, "*Meta v Bundeskartellamt*: Something Old, Something New" (2023) 8(3) *European Papers* 1077.

in individual professionals and cultivated over years of experience, can now be harvested through data and deployed in machines at scale. This long arc began with machines that mimicked the work of human hands; GenAI has brought the same dynamics to so-called "knowledge work" in a manner that has been startling for many. Consider the lawyer who uses some GenAI tool to help draft client correspondence. The tool may make her job easier, but it also learns from her practice – her way of framing arguments, her stylistic tics, her professional judgment. Something of her "aura" as a lawyer is captured, abstracted, and folded back into a system that can assist others or even replace parts of her role. Private law has little to say about this appropriation.

21 This illustrates a deeper issue: Data is not neutral. It is often the residue of personal labour, the trace of embodied knowledge or practical expertise. The rise of GenAI makes this increasingly visible, because the systems do not only "ingest" published works as spatio-temporal objects that can be broken down into so many "tokens" and digested by a model; they also appropriate style, voice, and patterns of reasoning that are implicit within the structure of those works. These are not easily fenced off as property, nor can they be adequately managed by contract or privacy law. They are forms of personal contribution that slip through the cracks of the liberal legal order.[26]

22 These dynamics play out on two related but *per se* very different registers. Firstly, there is the political economy register, which concerns things like workforce resilience, platform dominance, and technology's impacts on labour-capital relations. Labour-driven approaches to data have begun to grapple with this. Some proposals treat data as a form of work, requiring compensation or collective bargaining.[27] For example, Imanol Arrieta-Ibarra and co-authors have argued that in most accounts of the digital economy, user data is framed as a form of capital generated and owned by firms that monitor willing (contractual) counterparties. This framing erases the productive role of users themselves, leaving them without incentives, entrenching unequal distribution of value, and fuelling anxieties about automation. An alternative is to recognise data, at least in part, as a form of labour. Doing so, they say, could create a genuine market for user contributions and correct some of these distortions. Rebalancing

26 On the dignitarian aspect of this problem, see for example Salomé Viljoen, "Data as Property?", *Phenomenal World* (16 October 2020) <https://www.phenomenalworld.org/analysis/data-as-property/> (accessed 19 September 2025).

27 For discussion, see for example Julian David Jonker, "Is Data Labor? Two Conceptions of Work and the User-Platform Relationship" (2025) 35(2) *Business Ethics Quarterly* 153.

will require countervailing power – whether through competition policy, collective organisation of data labour, or regulatory intervention.[28]

23 Secondly, there is a very personal register. The author has had numerous conversations – none with a really definitive outcome – about what it is that "data protection" is trying to protect. There seems to be some connection between a person and "their" data – data that they generate and that others generate "about" them or in interaction with them – that is hard to define but deeply intuitive. Part of the problem, at least, is a kind of alienation that results when personal "traces" are appropriated without recognition, or for a purpose that is counter to the individual's interests. Perhaps a heightened sensibility for the personal dimension of data is needed: that it is not just an economic input, but also a fingerprint of human activity. Of all systems, liberal private law, grounded in autonomy and dignity, should be able to account for this – yet it does not. The gap is not lost on private lawyers. Much of the discussion in the Colloquium circled back to personality rights, publicity rights, and the tort of passing off as more promising avenues for development than copyright law. These bodies of doctrine at least recognise that something is at stake when aspects of a person's identity – name, likeness, reputation, or commercial *persona* – are appropriated without consent. They are fragmentary and unevenly developed across jurisdictions, but they point to a different orientation: one that treats the personal as carrying its own normative weight.

### *B. Beyond liberal private law: data as a "factor of production"*

24 In 2020, the State Council of the PRC formally recognised data as a new "factor of production", placing it alongside land, labour, capital and technology:[29]

> As a new factor of production (生产要素), data is the basis of digitalization, networkization (网络化), and intelligentization (智能化), and has been rapidly integrated into various aspects of production, distribution, circulation, consumption, and the management of social services, profoundly changing modes of production, living, and social governance. Construction of a basic

28 Imanol Arrieta-Ibarra *et al*, "Should We Treat Data as Labor? Moving Beyond 'Free'" (2018) 108 *AEA Papers and Proceedings* 38.

29 The Chinese Communist Party Central Committee & The State Council of the People's Republic of China, Opinion on Constructing a Basic System for Data and Putting Data Factors of Production to Better Use (2 December 2022) (translated by Centre for Security and Emerging Technology) <https://cset.georgetown.edu/publication/opinions-of-the-ccp-central-committee-and-the-state-council-on-constructing-a-basic-system-for-data-and-putting-data-factors-of-production-to-better-use/> (accessed 4 September 2025). The author has also used ChatGPT and Claude to translate the original text.

> system for data (数据基础制度) concerns the overall situation of national development and security. …

25 Subsequent policy initiatives, from the *Digital China* strategy to the establishment of national and provincial data exchanges, have reinforced this framing.[30] This marks an attempt to integrate data into the long-standing logic of economic planning in the Chinese communist paradigm: Identify key inputs, quantify them, and build institutions to optimise their allocation in a planned economy with pockets of free market enterprise. On this view, data is neither a private entitlement nor a by-product of digital life. It is a strategic resource, akin to energy or infrastructure. The state's role is to guarantee the orderly allocation of this factor, balancing efficiency with security through both market and non-market mechanisms.

26 The merits of this approach are clear on its own terms. Firstly, it squarely acknowledges that data is not trivial or peripheral but central to contemporary political economy. Secondly, it directs attention to the collective dimension of data: The value of a dataset lies less in individual entries than in its aggregation, such that institutional mechanisms are required to assemble, manage, and deploy it. Thirdly, it ties data governance explicitly to national strategy, recognising that the ability to control and harness data is a matter of economic sovereignty. In contrast to the fragmented and reactive posture of many liberal systems, the PRC framing is deliberate, systemic, and forward-looking. Yet, its productivist framing is also limited. It rightly contextualises data as a phenomenon of importance to the political economy. (The parallel narrative of "data as the new oil" mirrors the logic of a "factor of production" in the West.) But questions of personal autonomy or dignity are pushed to the periphery. By treating data purely as a factor of production, it strips away its connection to the persons who generate it.

27 Just as the author is concerned to critique liberal legal systems on their own terms, a Marxian vocabulary is useful for an internal critique of the data of production framing. The key to this critique is understanding Marx's Hegelian roots.[31] Hegel's account of labour emphasised not only

30 See, *eg*, The State Council of the People's Republic of China, "China Unveils Plan to Promote Digital Development" (28 February 2023) <https://english.www.gov.cn/policies/latestreleases/202302/28/content_WS63fd33a8c6d0a757729e752c.html> (accessed 4 September 2025). See also Lihua Huang *et al,* "Toward a Research Framework to Conceptualize Data As a Factor of Production: The Data Marketplace Perspective" (2021) 1 *Fundamental Research* 586.

31 The author's reading of Marx is influenced by Shlomo Avineri, *Karl Marx: Philosophy and Revolution* (Yale University Press, 2019), but the point is probably better developed in the earlier Shlomo Avineri, *The Social and Political Thought of Karl*

*(cont'd on the next page)*

production but the formation of selfhood through externalisation.[32] In Marx's thought, labour externalises human capacities into the world, but under capitalism, that externalisation becomes alienation in the sense of "estrangement" from oneself: the product of work stands apart from, and even against, the worker.[33] Ironically, the PRC's discourse exhibits a similar dynamic. Data, which is always in some sense a trace of human activity, is abstracted from its origin and recast as raw material for allocation. To ignore the subjective dimension of data is to ignore that it is a fingerprint of personhood, not just a fungible input for gross domestic product growth. Thus, by ignoring the *personal* dimension of data, the PRC approach suffers a blind spot of its own.

28 For liberal systems, the challenge is thus not to reject the PRC framing out of hand but to learn from what it makes visible. By naming data as a factor of production, the PRC underscores what liberal private law has failed to grasp: that data is central to an evolving political economy. We need to look at the *political* economy of data, not just the *economics* of data, because (for the reasons set out above) we cannot assume that liberal private law completely "depoliticises" market dynamics around data into neutral (*ie*, because consensual) market transactions. The reformist task for liberal private law is to hold both aspects together: to treat data as productive and allocatable, yet also as personal and dignitary. Here, fragments of doctrine such as personality rights, publicity rights, and passing off become newly salient. They show that liberal law does have tools to recognise when traces of identity and reputation are misappropriated. They are not comprehensive, but they embody a sensibility that could be extended. A liberal legal order need not – and should not – replicate the PRC's statist productivism. But it cannot continue to treat data as a legal non-entity.

29 If we step back from doctrinal detail and policy experiments, the core question is *institutional design*. How should we structure the legal order around data in a way that is fair, sustainable, and maximises

---

*Marx* (Cambridge University Press, 1968). See also Shlomo Avineri, "The Hegelian Origins of Marx's Political Thought" (1967) 21(1) *The Review of Metaphysics* 33. Avineri shows that Marx's concept of alienation remains central throughout his work: it is the estrangement of individuals from their capacities and traces of life, reconfigured as impersonal inputs. This insight helps frame today's data economy not just as exploitation, but as alienation – the appropriation of personal "auras" without recognition.

32 See generally Shlomo Avineri, *The Social and Political Thought of Karl Marx* (Cambridge University Press 1968), particularly at chs 1 and 4.

33 The German terms *Verfremdung* and *Veräusserung* are both often translated as "alienation"; the term *Verfremdung* which is the focus here, tracks most closely to modern English "estrangement" as reflected in the archaic and Scots English term "fremd".

social utility (including through stimulating AI innovation)? What is required is a liberal framework capable of recognising both the economic centrality and the personal aura of data. It is beyond the scope of this article to describe what that would have to look like. But it will, most likely, involve some recalibration of the familiar workhorses of liberal private law: contract and property. IP law, and copyright in particular, is likely to make an important cameo appearance at least.

## IV. To your positions!

30 The author was teaching "Ethics and Social Responsibility" at the time of writing, and had been preparing for a seminar on John Rawls's theory of distributive justice while working on this draft. Thinking through the Professor's Question and seeing how one's position in the data-driven economy so clearly impacts the debate around LLM data use, the author gained a new appreciation for Rawls's central thought experiment. As described by Timothy Hinton, Rawls's "original position" has been influential, *inter alia*, because it offers a compelling way of thinking about problems of justification and objectivity in political philosophy:[34]

> At the heart of these difficulties is the need to find an objective point of view from which to deliberate about matters of basic justice. Here 'objective' implies 'not mired in partiality' and 'not biased by one's particular position in the social world'. The original position is a hypothetical contractual situation in which parties who are ignorant about crucial features of themselves (such as how wealthy or talented they are, and what their vision of the best way to live is) are to select the principles of justice to regulate the basic institutions of their society. In selecting those principles, the parties are thought of as entering into an agreement that binds them to honor whichever principles they choose. By specifying that the parties are ignorant of matters that would allow them to favor themselves, Rawls vividly and unforgettably captures a widely shared sense that principles of justice cannot be justified by appealing to morally irrelevant considerations.

31 The analysis so far suggests two imperatives. Firstly, data should be recognised as an important fixture in the data-driven economy as it evolves – not incidental, not a by-product, and not neutral, but a factor that drives productivity and innovation and that has political implications and importance. Secondly, data should be recognised as personal – carrying with it traces of identity, expertise, and labour that should not be alienated without some appropriate and adequate *quid pro quo*. In Rawlsian terms, this means that principles governing data

34 Timothy Hinton, "Introduction: the Original Position and *The Original Position* – An Overview" in *The Original Position* (Timothy Hinton ed) (Cambridge University Press, 2015) at p 1.

must be consistent with the basic liberties of individuals as free and equal agents, while also structuring incentives so that the benefits of innovation are widely shared.

32 Recall the Professor's Question: "Wouldn't we have designed a better legal regime for LLM model developers' use of copyright material if we had known what we know now?" From this perspective, familiar moves like copyright minimalism or contract adhesion would quickly be ruled out. No one in the Original Position would choose a system in which, by bad luck, they might be born an ordinary user whose data is endlessly harvested and commodified without meaningful recourse, let alone an author or artist whose style is appropriated and copied by machines. Equally, no one would choose a system that treated all data as sacrosanct private property, stifling innovation and freezing technological progress to the general detriment. Behind the veil, the rational choice is one that secures individual dignity and autonomy, but also permits flows of data under conditions of reciprocity and fair return to create a functional political economy. The question then becomes: How would such a middle path look in a liberal legal order?

33 The question whether the law should be changed to *create* a licensing regime for model training must not be derailed by arguments that the *current* law of copyright may not, and in many cases does not, provide a legal basis for such a right because fair use (and other doctrines) permit non-infringing non-consensual use. Such arguments do not address the actual point in issue. Rawls helpfully brings the focus back: Putting respective interests aside, would any reasonable person agree that authors, artists, experts, *etc*, should have no right over their works that can be used to capture *any kind of rent* even as those works are being aggregated and used to create machines that can *directly compete with them*? This is a question that is at once determinative of very personal interests, and also crucial to the success of today's digital political economy.

### *A. The flawed premise of Web 2.0*

34 To answer the Professor's Question properly, the author proposes a brief survey of the journey taken so far. The Web 2.0 era – roughly the mid-2000s to the early 2020s, bookending the arc of social media's first decade – was premised on the idea that Internet users would produce content as well as consume it. For a few years, this was exciting and dynamic. And then the rot set in. It is no accident that "enshittification"

was the Macquarie Dictionary's word of the year in 2024.[35] As Cory Doctorow, who coined the term, explains: "First, platforms are good to their users; then they abuse their users to make things better for their business customers; finally, they abuse those business customers to claw back all the value for themselves."[36]

35 Web 2.0 arrived with the promise of a more democratic internet. Platforms invited users to share their voices, their creativity, and their communities. Blogs, wikis, social networks, and video sites removed barriers to publication and distribution. Yet, the apparent openness of this architecture concealed an extractive design: The price of participation was enclosure within proprietary platforms, whose business models were premised on harvesting data. When a user interacts with an application owner, the latter gets all the user's data, subject only to the terms of an adhesive contract and some data protection regimes that developed after the fact. The fatal flaw was thus an architectural feature of the Internet itself, that we probably would have designed differently *had we but known*. Instead of contributing to a digital commons, users found themselves producing content for intermediaries who accumulated power and profits.

36 First, they used that data to maximise user attention on the platform, notably through targeted advertising. Over time, the incentive structures implicit in this dynamic gave rise to progressive platform self-optimisation at the expense of users and complementors. Content that was once valued for originality or community resonance became calibrated to click-through rates, algorithmic visibility, and advertising yield. The attention economy rewarded speed, controversy, and catchiness, while sidelining slower forms of knowledge and expression. At the political-economic level, this meant not only a degradation of content quality but also a consolidation of market power: Platforms that controlled distribution controlled value, and the reciprocity of contribution was steadily eroded. In a world of infinite content, he who puts content in front of eyeballs is king.[37]

37 The rise of GenAI illustrates a new phase of this same logic, in which data is monetised in a new way. The enormous archives of

35 Tony Shepherd, "'What Many of Us Feel': Why 'Enshittification' Is Macquarie Dictionary's Word of the Year", *The Guardian* (26 November 2024) <https://www.theguardian.com/science/2024/nov/26/enshittification-macquarie-dictionary-word-of-the-year-explained> (accessed 9 October 2025).

36 See Cory Doctorow, "Tiktok's Enshittification", *Pluralistic* (21 January 2023) <https://pluralistic.net/2023/01/21/potemkin-ai/#hey-guys> (accessed 9 October 2025).

37 While the words are the author's own, the idea is from Georg Zoeller of the Centre for AI Leadership, Singapore.

user-generated content accumulated by platforms – blog posts, forum threads, photos, videos, comments – are now being repurposed as training data for large-scale models. This represents a double extraction: First, user contributions are monetised indirectly through advertising, with little or no return to the creators themselves. Now those same contributions are being mined once more to build systems capable of generating synthetic substitutes for the very kinds of work users produced. The free labour of Web 2.0 has become the raw material for a new wave of automation, often without consent, compensation, or even acknowledgement.

38 Seen in this light, the failures of Web 2.0 are precursors to today's debates about generative AI. The same structural features – platform dominance, network effects, and extractive business models – shape the trajectory of both. If Web 2.0 hollowed out the ecology of online content by privileging attention over substance, generative AI threatens to hollow out the ecology of cultural production more broadly by turning yesterday's contributions into tomorrow's synthetic competitors. In the French classic *Le dîner de cons*, invitees are flattered to be invited to a fancy meal, but it all goes wrong when they realise they are the entertainment. The Internet rule of thumb runs that if it is free, you are the product. In the third stage of enshittification, you are (*ie*, your data is) the product – often, even when you are a paying customer. The challenge for law and policy is to break this cycle of enshittification; if not, it will repeat at higher levels of technological abstraction.

### *B. The failed promise of Web3*

39 As Web 2.0 descended into enclosure, Web3 was billed as the cure. Its advocates promised disintermediation, decentralisation, and new forms of ownership. No longer would users be trapped within walled gardens or beholden to platform monopolies. Instead, blockchain-based protocols would allow individuals to transact directly, with tokens aligning incentives and distributing value more fairly across the network. On paper, this was an attractive vision. It acknowledged many of the failures of Web 2.0 and sought to design them out to give users a stake in the digital economy they helped create. Yet in practice, the results fell far short.

40 Non-fungible tokens ("NFTs"), for example, were hailed as tools for anchoring authenticity and provenance on-chain, and the author recalls many conversations with digital enthusiasts about new ways to capture value from their work through minting NFTs. However, their use has been overwhelmingly speculative rather than infrastructural. The NFT bubble burst in 2022, but the author is not aware of any serious attempt to deploy NFTs to solve the problems raised by the advent of generative

AI – for example, as a layer to distinguish original human output from AI or synthetic material or to preserve attribution or personality rights. Again, the infrastructural aspect of the problem – how to distinguish original from synthetic, or how to preserve attribution across remix and reuse – remains unsolved. The rhetoric of decentralisation also proved hollow. Far from eliminating intermediaries, Web3 produced new ones: exchanges, marketplaces, token-gated platforms, each with their own asymmetries and rent-extraction. "Community governance" often boiled down to concentrated token holdings or informal hierarchies, replicating many of the same pathologies as Web 2.0. What was meant to be a protocol-driven commons became, for the most part, a new theatre of speculation.

41 The same pattern is evident in data governance. If Web3 were genuinely committed to breaking platform dominance, one might have expected more serious calls for a redesign of the Web's basic architecture – a system in which individuals could interact with services without surrendering all their behavioural and creative data. There are noteworthy counter-examples. "Solid" (for SOcial LInked Data) is a platform for linked-data applications that are completely decentralised and fully under users' control rather than controlled by other entities.[38] It is, in fact, an effort to repair what is broken in the World Wide Web by one of the Web's chief architects, Sir Tim Berners-Lee. But Solid is a "Web3" project in the broader sense that has no inherent connection to blockchain technology. Further, with great respect to Sir Tim and his collaborators, we are a long way from turning back the tide, and the author doubts whether it is a realistic goal at this stage. The best possible outcome might be to splinter off a new, fairer Web; but that undermines the very premise of a single Web and would likely track current geopolitical fault-lines in cyberspace.[39]

42 Another noteworthy exception – this time with blockchain and "tokenomic" DNA – is the Basic Attention Token ("BAT"),[40] tied to the Brave browser.[41] Whatever its limitations, BAT tries to realign incentives by rewarding users for their attention and redistributing value in a more transparent way. Unlike the bulk of crypto projects, which leaned on speculative trading or vague appeals to "community", BAT started

38 This project was initiated by Sir Tim Berners-Lee (of World Wide Web fame) at Massachusetts Institute of Technology; for a time it was run under a company, Inrupt, and is now under the stewardship of the Open Data Institute. See further <https://solidproject.org> (accessed 3 September 2025).

39 See, *eg*, Kieron O'Hara & Wendy Hall, *Four Internets: Data, Geopolitics, and the Governance of Cyberspace* (Oxford University Press, 2021).

40 See further <https://basicattentiontoken.org> (accessed 3 September 2025).

41 See further <https://brave.com/brave-rewards/> (accessed 3 September 2025).

with a plausible diagnosis of the distortions of the attention economy and offered a *protocol-level* fix. It demonstrated that tokens could serve as instruments for restructuring the economics of the web. Yet notice what has not emerged, even in the era of GenAI: a Basic Content Token ("BCT"). If attention could be tokenised and rewarded, why not creative contribution itself? A BCT would allocate value not just at the point of consumption but also at the point of production – recognising that content is the generative force of the digital ecosystem. Each post, image, or comment could trigger a micropayment, however small, ensuring that value flowed back to creators rather than being siphoned off by platforms. Such a system would not solve all the problems of Web 2.0, but it would have realigned incentives toward quality and sustainability rather than clickbait and extraction.

43 The absence of a BCT reveals the limits of Web3 as it was actually pursued. Instead of confronting the architectural substrate of the digital economy, the movement largely contented itself with financialisation at the edges: NFTs as speculative assets, decentralisation as slogan. Deeper reforms – such as watermarking human creativity, user-controlled data channels, or universal protocols for content remuneration – are still aspirational. In hindsight, this was the real missed opportunity: to take the lessons of Web 2.0's enshittification and design a sturdier foundation before GenAI developers – themselves mostly large data companies – came along to mine the ruins.

44 Generative AI has the potential to turbocharge human creativity, enhancing our efforts and synthesising our "special sauce" with the stock of ideas and expressions that have gone before to create novel combinations. But the first generation of models has been built on data either appropriated without consent, misappropriated, or literally stolen – from Meta using books downloaded from pirate torrents, to the ubiquitous use of Common Crawl, to the "models-all-the-way-down" problem.[42] David Carson writes in strident terms:[43]

> One of AI's original sins (there are many) is scraping every corner of the Internet for training data, sucking up text, photos, video, music and anything they can find online to feed the large language models tech companies are racing to develop. Intellectual property rights and copyrights be damned, in true Silicon Valley style, AI companies are moving fast and breaking things. They need the data, so they take it and claim it's transformative fair use.

42 Christo Buschek & Jer Throp, "Models All the Way Down", *Knowing Machines* <https://knowingmachines.org/models-all-the-way#section2> (accessed 3 September 2025).

43 David Carson, "Theft is Not Fair Use", *Medium* (21 April 2025) <https://jskfellows.stanford.edu/theft-is-not-fair-use-474e11f0d063> (accessed 3 September 2025).

45 Had this been thought about in advance – with knowledge of what is now known from Web 2.0's enclosure, Web3's speculation, and GenAI's appropriation – it is clear that a more plausible licensing regime for training data would have been built. The horse has bolted, it is true. Existing foundation models cannot be dismantled, and it is entirely possible that the law *as it existed at the time* will not yield a future-proof resolution of the underlying distributive problems.

### *C. Adrift on a rising tide of slop …*

46 But the target is a moving one. The trajectory from Web 2.0 through Web3 has led to the present moment, where GenAI systems feed on the accumulated detritus of two decades of user-generated content and hoover up copyright material in a legal grey-zone. But what happens when the well itself begins to run dry, or worse, becomes poisoned? Over the past year, concern has grown about the sheer volume of AI-generated material on the public Internet. The term "AI slop" has entered circulation to describe this rising tide of synthetic text, images, audio, and video. Slop is not just poor-quality content. It destabilises the data ecosystem itself.[44] Firstly, it creates incentives to remove content from the public internet and put it behind a paywall: Why generate valuable content if users read the AI-generated summary rather than visiting the source website (and viewing its ad banners) at all? Secondly, and more importantly, when synthetic material re-enters the training pipeline, models begin to reproduce artefacts of earlier generations, amplify distortions, and converge on a statistical mean. Diversity shrinks, fidelity drops, and outputs lose texture. Researchers call this "model collapse": The ecosystem starts to cannibalise itself.[45] From what the author understands, there is no reliable technical fix. Filtering requires reliable detection, but AI-generated content is often indistinguishable from human work. Watermarking schemes, such as those now mandated by law in the PRC, may help at the margins, but they are patchy, easy to strip, and difficult to enforce.[46]

---

44 See, *eg*, Emma Marris, "AI Slop Might Finally Cure Our Internet Addiction", *The Atlantic* (22 July 2025) <https://www.theatlantic.com/technology/archive/2025/07/ai-slop-internet-addiction/683619/ (accessed 4 September 2025). See also the unpublished but very scholarly study by Jane Tullis, "Sifting Through the Slop: How Generative AI Created a Market for Lemons for Text-Based Works" (1 April 2025) <https://ssrn.com/abstract=5266660> (accessed 4 September 2025).

45 See, *eg*, Ilia Shumailov *et al*, "AI Models Collapse When Trained on Recursively Generated Data" (2024) 631 *Nature* 755.

46 See Coco Feng, "China's Social Media Platforms Rush to Abide By AI-generated Content Labelling Law", *South China Morning Post* (1 September 2025) <http://scmp.com/tech/policy/article/3323959/chinas-social-media-platforms-rush-abide-ai-generated-content-labelling-law> (accessed 4 September 2025).

47 We are adrift on a rising tide of AI-generated slop, and, absent some "Hail Mary" breakthrough, the marginal cost of identifying authentic data will rise sharply as the tide of slop swells. The greater the volume of slop, the more difficult it becomes to find novel inputs for the next generation of training data. Human-authored contributions – long treated as free raw material – are thus *becoming scarcer*. If the scenario does not change radically, then "genuine" human output will only become more valuable. Some practitioner and technical colleagues have already suggested, in our conversations, that model developers may in future *commission original works* simply to ensure clean training input. This is the target for law reform efforts. However, it is uncertain whether the context or problem is always accurately understood, even in the legal community. A related discussion in our Colloquium was whether Singapore would have adopted its liberal text and data mining exception in 2021 had the implications of GenAI been on the horizon.[47] A consultation is currently underway on this point and it seems, despite concerns raised by key stakeholders, that further expansion may in fact be on the table.[48] In other words, the legal framework for markets in training data is still in flux, but it is unclear which way things are going to go.

## V. Back to the spaceship!

48 Imagine a spaceship, travelling to a destination unknown and deciding the rules of the game to govern this new society. Behind the veil, reasonable persons would want two things, as noticed above: They would want (a) rules that distribute benefits and burdens *fairly*. This would mean reasonable opportunities (supported by appropriate legal and market structures) for collecting fair rents and for preventing competition by machines trained on their data footprint. Fairness cuts both ways – they would not want these regimes to be too onerous, such that innovation were hampered, but it is highly unlikely they would collectively agree to the open-slather approach we have, in fact, taken in the first decades of the data-driven economy. Following on from this, reasonable persons would want (b) rules that sustain a healthy, *functional* data ecosystem, rather than one liable to collapse (*eg*, through inundation with slop).

47 See Ng-Loy Wee Loon, "Copyright Exceptions for Text and Data Mining: A Case of Specificity (Certainty) and Generality (Flexibility)" in *Kreation Innovation Märkte – Creation Innovation Markets: Festschrift Reto M Hilty* (Florent Thouvenin *et al* eds) (Springer, 2024).

48 See Pin-Ping Oh, "Potential Expansion of Singapore's TDM Exception?", *Bird & Bird Insights* (26 April 2024) <https://www.twobirds.com/en/insights/2024/singapore/potential-expansion-to-singapores-tdm-exception> (accessed 4 September 2025).

49 This is why the Professor's Question – whether we ought to have designed a licensing system for training data from the outset – must be taken seriously *going forward*. It may be too late to claw back what has already been appropriated to build the first generation of models. But for the next wave, where human traces become the scarce premium input, the legal order cannot afford to make the same mistake again. The spaceship has not landed; in fact, it will never land – it is just orbiting the sun. Its journey is not through space, but time, and in every generation, society bears the responsibility of making decisions that will affect those who come after.

50 What, then, is to be done? Lawyers play a lead role in crafting a viable data-driven economy for the future.[49] Katharina Pistor's pithy concept of the "code of capitalism" is a useful reminder that markets are not natural phenomena; they are coded into existence by law.[50] Accordingly, law is not something exogenous to economics; it is infrastructural, providing the scaffolding of enforceable contracts and property rights, as well as the outer limits of what is permitted within the four corners of the game. Property, contract, corporate form, and finance act as "modules" that structure entitlements, allocate risk, and channel rents. Financial assets exist and circulate only because the law endows them with recognisable attributes. Tortious, criminal, anti-competitive, and deceptive conduct, for example, is out not (only) because it is immoral, but because it undermines the *very idea of a market* as a rule-bounded context in which economic agents make meaningful choices about economic transactions, and so leads to market failures.

51 When law codes markets badly – including by *failing to code* something important – it produces not liberty and exchange but failure and collapse. Data's treatment as an unregulated, unprotected commons is not inevitable. The architecture of extraction (and over-extraction) is a legal artefact as much as it is an accident of technical design. We somehow failed to appreciate the role that data would play in the digital economy as it evolved, and so the technical architectures we built were simply not designed to stop data appropriation at hyperscale. Slop dramatises this: It is the visible symptom of a market coded for extraction without reciprocity and innovation without sustainability. By failing to build legal categories that recognise the dual character of data – as both

49 See Katharina Pistor, "A Legal Theory of Finance" (2013) 41(2) *Journal of Comparative Economics* 315 for a crisp presentation of the central idea that financial markets are legally constructed.

50 See Katharina Pistor, *The Code of Capital: How the Law Creates Wealth and Inequality* (Princeton University Press, 2019).

productive input and personal trace – liberal systems effectively licensed the degradation of the infosphere. We are now living with the results.

### *A. A six-point plan*

52 A reform agenda for the data-driven economy cannot be limited to patching copyright at the edges or tinkering with adhesion contracts. It must be programmatic, recalibrating the workhorses of liberal private law while also experimenting with new institutional forms. The aim is not to mimic the PRC's productivist model but to avoid liberal law's recurring failure to recognise both the *productive* and the *personal* dimensions of data, and the modern tendency to isolate "economics" from the realm of the political. The following items are important talking points that should be on the agenda. These are drawn from the author's own research but are informed to a great extent by the Colloquium. They are intended to provide a useful framework for others, regardless of the position they occupy in the debate.

53 The first step is to *recalibrate contract and property law*. At present, most data exchanges are structured through unilateral adhesion contracts that enforce platform control while disclaiming any recognition of user contributions. What is needed is a framework that treats human-generated data not as "exhaust" but as a valuable input, entitled to recognition and remuneration. This does not necessarily mean conferring absolute property rights – which could well stifle innovation – but it could well involve creating quasi-proprietary or labour-like entitlements that prevent unilateral appropriation and allow for fair returns. These should also fit coherently with quasi-proprietarian and regulatory rights that already exist in data protection (and similar) regimes.

54 Secondly, we need to *develop licensing and collective rights mechanisms fit for the age of GenAI*. The current patchwork of addenda and opt-out forms is not a market, but a stopgap. *Inter alia*, what would make the market more functional is collective licensing: societies or *consortia* that can negotiate on behalf of creators, pool rights, and distribute returns. This is how music copyright was civilised into something workable, and there is no reason why analogous mechanisms could not be built for training data. Critically, these mechanisms could extend beyond copyrightable "works" to cover *styles, traces, and expertise* – the qualities that generative models learn and replicate, but which current doctrines do not protect.

55 Thirdly, moving closer to IP *per se*, we should *revisit personality and publicity rights*. These doctrines are often treated as marginal, but they directly address the appropriation of voice, likeness, and identity.

They could be expanded to cover not only obvious cases of deepfakes or celebrity exploitation, but also the subtler forms of style and *persona* appropriation that underpin much of generative training. This is one way liberal law can give effect to the dignitary dimension of data without collapsing into absolutist property claims.

56 Fourthly, the *integrity of the data commons* must be treated as a public good. "Digital law" should impose obligations to prevent the collapse of the infosphere. This could include provenance-tracking requirements, watermarking mandates, and liability rules for synthetic content pollution. The goal is not to eliminate AI-generated material, but to ensure that authentic human contributions remain visible, attributable, and usable for future innovation.

57 Fifthly, we should be open to *protocol-level solutions*. The philosophy behind the Solid project is compelling: the architecture of our communications protocols should reflect our normative ideals and values. This could extend from data ownership by design to micropayments, for example in the service of individual or collective licensing arrangements. Web3 may have disappointed, but its more thoughtful experiments point toward what might be possible. The BAT showed how tokens could be used to restructure incentives around advertising. Something like a BCT could do the same for creative contributions. Imagine a system where each post, image, or article generated a micropayment at the protocol level, ensuring that value flowed back to contributors rather than being siphoned off by intermediaries. Such systems should not be oversold – but they deserve serious testing, especially in verticals like publishing, visual arts, and education, where the extraction problem is most acute.

58 Finally, there is a need for *institutional pluralism*. No single reform will be sufficient. Some problems are best addressed through private law recalibration; others through public regulation of platforms; still others through technological or protocol-level experiments. The danger is in assuming that the current mix can carry the weight of a data-driven political economy. It cannot.

## VI. Conclusion

59 The path ahead will require imagination, but also discipline. The promise of liberal private law is a normatively robust claim to balance freedom with structure, autonomy with reciprocity. Those values are still fit for purpose, but only if we are willing to update, from time to time, the doctrines through which they are realised. GenAI has revealed a blind-spot in liberal private law: Data is a central input of the digital

political economy. (The author has not considered other inputs, like compute, transmission infrastructure, or energy, at this point.) Data is not just "exhaust" produced by our interactions with online services and environments, but the stuff from which value is created. The rise of AI slop gives the underlying problem present currency. Human traces – text, images, voices, gestures – are becoming the scarce factor in a polluted digital ecosystem. The Professor's Question resounds with amplified force.

60 The first generation of models may have been fuelled by misappropriated content, but the path ahead is not fixed. Rawls's perspective is a reminder that these decisions affect not only those living in the present but also future generations. According to Pistor, the rules society chooses will literally code the market. The reform agenda to be drafted will almost certainly include some elements of IP, and copyright in particular. But it must be plural. It also includes recalibrating contract and property to prevent unilateral enclosure, building collective licensing mechanisms to distribute rents, expanding personality-based rights to respect identity and style, safeguarding the data commons against collapse, and experimenting with protocol-level interventions. The latter might include radically rethinking the assumptions of the Internet's basic architecture, such that users retain a greater measure of control over their data. Alternative approaches to data, such as the PRC factor of production framework, provide useful counterpoints that can stimulate an internal reformist agenda. The challenge is to ensure that the legal architecture of the digital economy is fit for the future, not simply a repetition of past mistakes. If society accepts that human contribution will be the scarce premium in an age of synthetic churn, then the task is clear: institutions must be designed to value and preserve it.

## VII. Postscript

61 The great irony of writing this paper is that the author used GPT-4o and GPT-5 more closely in the writing process than ever before. To be clear: ChatGPT was not asked to provide ideas or used to replace research. A different problem was encountered – there were three sprawling drafts, each pursuing a different line of enquiry that had arisen over the 24 months since the start of the Colloquium, each getting a little lost (to be honest) in the weeds. There was very limited time to bring them into a coherent whole, and hence, the bot was used primarily to parse the author's work and help bridge ideas partially developed in each draft. Thus, this paper is both the result of "slow science" – *ie*, a drawn-out Colloquium, other events and conversations, over a long period – and of a last-minute effort to smash out a piece of coherent writing before a journal deadline, during a teaching term.

62 In one sense, it is a good example of the value this kind of tool can add to academic work. The author is happy with the product and has no academic or ethical concerns about the process involved. However, *having more time to write this paper would have been preferred*. Behind personal dissatisfaction with writing up "slow thinking" in a last-minute rush is a deeper worry. That worry is twofold. Firstly, the availability of these tools will encourage people to write articles that are *not* the product of slow thinking. Already, scientific publishing is drowning in a wave of slop, and the managerialist, metric-driven direction of university and other managers around the world is likely to interact with GenAI in counterproductive ways. Where chatbots can mimic reasonably competent scholarly output in seconds, it behoves us to think very carefully about how we work, and what value human creativity and effort has in such a knowledge ecosystem. Secondly, the author worries that the availability of these tools will encourage scholars to maintain a cadence of work that is, frankly, inhuman – encouraging them to crowd their calendars, as the author's is currently crowded, in reliance on the availability of *deus ex machina* to help finish the last act. For both reasons, the author is glad that he began his career in the pre-GenAI era and is equally concerned for, and about, those coming up now. This is a sentiment that surely resonates with others.

---