DO YOU COPY?: ATTRIBUTING COPYRIGHT INFRINGEMENT TO ACTORS INVOLVED IN TEXT-TO-IMAGE GENERATIVE ARTIFICIAL INTELLIGENCE

AMY WOJCIECHOWSKI* AND DANIELA SIMONE**

Generative artificial intelligence ('GenAI') models democratise the artistic landscape by enabling amateurs and artists alike to create quality images. At the same time, copyright holders are increasingly concerned about copyright infringement where their works have been incorporated into training data without authorisation. We undertake a step-by-step analysis informed by technical information from recent scholarship and overseas litigation to consider the creation of training data, its use in machine learning, then the development and deployment of text-to-image GenAI models. This 'supply chain' approach allows us to assess the liability of actors at each stage, revealing the Australian legal position to be significantly more complex than has been appreciated. We explore the implications of this analysis, arguing in favour of a principled, rather than pragmatic, application of legal principles. We conclude by recommending reform agendas be animated by a broad view of copyright's aims as well as an appreciation of its limits.

I INTRODUCTION

Shortly after the public release of ChatGPT in 2022, generative artificial intelligence ('GenAI') technologies exploded into the public consciousness. These technologies promise to revolutionise the ways we work and create, but there is uncertainty about their copyright implications. In the face of ongoing global litigation,[1] copyright law has been seen as a potentially significant brake on

* BCom/LLB (Hons 1), Macquarie University.

** Senior Lecturer, Macquarie University; Honorary Lecturer, University College London.

We would like to thank Professor Kimberlee Weatherall and the anonymous reviewers for their very helpful feedback. Thanks also to Associate Professor Rita Matulionyte for comments on an earlier version submitted as Amy Wojciechowski's Honours thesis. Any errors or omissions remain our own.

1 Professor Edward Lee of Santa Clara University School of Law maintains a map of current Copyright/AI lawsuits in the United States ('US') on his blog, *ChatGPT Is Eating the World.* As at 17 May 2025 there were 41 ongoing cases: Edward Lee, 'Latest Map of All 41 Lawsuits v AI Companies (May 17, 2025)', *ChatGPT Is Eating the World* (Blog Post, 17 May 2025) <https://chatgptiseatingtheworld.com/2025/05/17/latest-map-of-all-41-lawsuits-v-ai-companies-may-17-2025>. For a map of global disputes, see Edward Lee, 'World Map of Copyright Lawsuits v AI (Mar 29, 2025)', *ChatGPT Is Eating*

the development and use of GenAI.[2] Some cases seem to represent a David and Goliath battle, positioning multinational technology companies such as Microsoft and NVIDIA as defendants and individual artists as plaintiffs. Despite over 40 legal cases having been brought before overseas courts, including in the United States ('US'),[3] United Kingdom ('UK'),[4] and the European Union ('EU'),[5] it is likely that many of the answers to the important legal questions raised by plaintiffs remain a way off. It is unknown how an Australian court might approach the copyright infringement questions.[6] Although litigation has not yet been instigated in Australia, the challenges artificial intelligence ('AI') poses to copyright are widely discussed in legal and academic circles. The potential need for reform to copyright law has been on the Federal Government's radar.[7]

This article takes a closer look at whether, and how, liability for copyright infringement could be attributed to the various actors involved at each stage of the text-to-image GenAI process in Australian copyright law. It reveals that the position is significantly more complex than previous scholarship admits. Limits of space make it necessary to confine our focus to text-to-image generation models, but many of our insights are more broadly applicable. This focus allows us to delve into the technical and doctrinal complexities in greater detail than would otherwise be possible within the scope of a journal article. Text-to-image models highlight some of the nuances of the technology and its advances. They have been some of the first instances to generate litigation and provide a vivid illustration of some of the messier copyright issues that have concerned right holders, such as the

---

the World (Blog Post, 29 March 2025) <https://chatgptiseatingtheworld.com/2025/03/29/world-map-of-copyright-lawsuits-v-ai-mar-29-2025/>.

2 Carys J Craig, 'The AI-Copyright Trap' (2025) 100(1) *Chicago-Kent Law Review* 107.

3 See, eg, *Getty Images (US) Inc v Stability AI Inc* (D Del, No 1:23-CV-00135-UNA, 3 February 2023) ('*Getty (US) v Stability AI* Complaint'); *The New York Times Company v Microsoft Corporation*, 777 F Supp 3d 283 (SD NY, 2025) ('*New York Times v Microsoft*').

4 *Getty Images (US) Inc v Stability AI Ltd* [2024] FSR 12. See also, *Getty Images (US) Inc v Stability AI Ltd* [2025] EWHC 2863 (Ch) ('*Getty Images 2025*'), handed down just before publication of this article.

5 *Robert Kneschke v LAION eV*, Landgericht Hamburg [Hamburg Regional Court], 310 O 227/23, 27 September 2024 ('*Kneschke*').

6 The issue may also be of broader relevance if proposed AI guardrails were to come into effect in Australia. Guardrail 3 requires that training data must be 'legally obtained': Department of Industry, Science and Resources, Australian Government, *Safe and Responsible AI in Australia*: *Proposals Paper for Introducing Mandatory Guardrails in High-Risk Settings* (Proposals Paper, September 2024) 37. Note that the proposed guardrails would only be mandatory in 'high–risk' cases, and voluntary in other circumstances. Questioning what this would entail from a copyright perspective, see Rita Matulionyte, 'Australia's Proposed Guardrails for High-Risk AI and Copyright Law', *Kluwer Copyright Blog* (Blog Post, 2 December 2024) <https://copyrightblog.kluweriplaw.com/2024/12/02/australias-proposed-guardrails-for-high-risk-ai-and-copyright-law>.

7 Understandably, issues of AI safety and the regulation of high-risk implementations may be higher up the list of priorities than copyright issues. Acknowledgment of the potential need for copyright reform can be seen in the appointment and activities of the Copyright and AI Reference Group: 'Copyright and Artificial Intelligence Reference Group (CAIRG)', *Attorney-General's Department* (Web Page) <https://www.ag.gov.au/rights-and-protections/copyright/copyright-and-artificial-intelligence-reference-group-cairg>. Most recently the Productivity Commission has suggested the need for copyright reform in response to issues created by AI: Productivity Commission, *Harnessing Data and Digital Technology* (Interim Report, August 2025) 24–8 <https://www.pc.gov.au/inquiries/current/data-digital/interim/data-digital-interim.pdf> ('*Productivity Commission Report*').

appropriation of an artistic style. As will be further explained, and as the example of text-to-image GenAI demonstrates, answers to some of the key infringement questions turn on specific technical mechanisms that operate within a particular model. They also depend on a court's approach to applying established legal principles to new circumstances.

Until recently, it has been assumed that the training and use of GenAI models raise relatively straightforward copyright infringement issues, with commentary focusing on the adequacy of exceptions in this context.[8] As more technical details about the GenAI supply chain come to light, and overseas litigation airs new infringement theories, the time is ripe for a more detailed consideration of the application of Australian copyright infringement provisions than appears in existing literature.[9] Synthesising new technical understandings about the GenAI 'supply chain' within a broader view of the interpretative options available allows us to provide a more nuanced assessment of the legal position in Australia than is commonly admitted.[10] A clearer view of the interpretative approaches available within copyright law provides essential information for stakeholders, advocates, and policymakers.[11]

This article begins by setting out the background to the issue and situating our contribution within the existing literature. Part III considers liability for copyright infringement in a step-by-step process across each stage of the GenAI supply chain: from the creation of data training sets; to their use in machine learning; to the development and fine-tuning of foundational AI models; and then the implementation of those models to create specific outputs. Part IV reflects on the implications of this doctrinal analysis. When considering how copyright should respond to GenAI challenges, we stress the importance of maintaining conceptual coherence and adopting a broader normative lens that understands that the role of copyright in regulating human creativity and communication extends beyond incentives. We conclude by noting limits to copyright law's capacity to provide the answers that creators and the creative industries may seek in this area.

---

8 For example, Rita Matulionyte, 'Australian Copyright Law Impedes the Development of Artificial Intelligence: What Are the Options?' (2021) 52(4) *International Review of Intellectual Property and Competition Law* 417 <https://doi.org/10.1007/s40319-021-01039-9> ('What Are the Options?'). A similar approach is observable in much of the international scholarship and debate.

9 Technical knowledge is sure to continue to advance and more may be revealed in the ongoing litigation: see above nn 1, 3–5. We rely on information available at the time of publication located within the best efforts of the authors as of 31 May 2025.

10 Katherine Lee, A Feder Cooper and James Grimmelmann, 'Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain' (2025) 72(2) *Journal of the Copyright Society of the USA* 251.

11 This analysis can shed new light on the assumed need for particular reforms to enable the development and deployment of GenAI, as well assumptions that these activities are straightforwardly infringing where unlicensed training data is involved.

II BACKGROUND

The social and economic impacts of text-to-image GenAI are no longer hypothetical and are becoming more visible in the Australian context. The rapidly improving capabilities of text-to-image GenAI models portend significant disruption in the creative industries.[12] The efficiency and quality of the images output by these models have exponentially increased such that artists are beginning to view the technology as a genuine competitive threat.[13] Users can guide GenAI to mimic an artist's style or previous work, leading to questions of copyright infringement based on the image output.[14] Some up-and-coming artists are concerned about sharing their work online due to the potential 'scraping' of their artworks to train the AI models that could replace them.[15] A prominent example is Polish digital artist Greg Rutkowski, who publicly expressed concern at how well an image-generating AI model could replicate the well-known fantasy style of his artworks.[16] Indeed, it emerged in late 2022 that users of image-generating AI models were embedding the prompt 'Greg Rutkowski' far more than even 'Picasso'.[17] With such success in replicating a signature style, Rutkowski's concerns echo those of many artists who have questioned whether their artworks were scraped from the internet to train GenAI models without their authorisation (or even knowledge). But the picture is nuanced. Some artists have been more positive about the technology and are actively embracing it in their practice.[18]

Creative industry concerns are exacerbated by uncertainty in the domestic copyright framework. The process of creating and using GenAI models seems likely to have involved the unauthorised use of copyright protected works, raising issues as to potential infringement of artists' and copyright holders' rights. The legal uncertainty surrounding the implications of GenAI's use of protected works

---

12 Attorney-General's Department, Australian Government, 'Artificial Intelligence (AI) and Copyright' (Outcomes Paper, Roundtable on Copyright, 18 December 2023) ('Outcomes Paper on AI and Copyright'); Select Committee on Adopting Artificial Intelligence, Parliament of Australia, *Final Report* (Report, November 2024) ('*Senate Report*').

13 Mark A Lemley, 'How Generative AI Turns Copyright Upside Down' (2024) 25(2) *Columbia Science and Technology Law Review* 21 <https://doi.org/10.52214/stlr.v25i2.12761>.

14 Michael D Murray, 'Generative AI Art: Copyright Infringement and Fair Use' (2023) 26(2) *SMU Science and Technology Law Review* 259 <https://doi.org/10.25172/smustlr.26.2.4>.

15 Ibid.

16 Melissa Heikkilä, 'This Artist is Dominating AI-Generated Art. And He's Not Happy About It', *MIT Technology Review* (online, 16 September 2022) <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/>.

17 Ibid. Market interference and the substitutability of GenAI outputs are key creator worries. In some cases, there may be concern about reputational damage occasioned by the circulation of works that imitate a creator's distinctive style. It is beyond the scope of this article to consider the consumer law, passing off, or moral rights issues that might arise in such circumstances.

18 On the opportunities for the creative industries: see *Senate Report* (n 12) [4.55]–[4.61]. Some suggest the general sentiment against AI is shifting: Heikkilä (n 16); Rudi Zygadlo, 'Should Artists Be Terrified of AI Replacing Them?', *The Guardian* (online, 11 August 2024) <https://www.theguardian.com/lifeandstyle/article/2024/aug/11/should-artists-be-terrified-of-ai-replacing-them>; Claudia Baxter, 'AI Art: The End of Creativity or the Start of a New Movement?', *BBC News* (online, 21 October 2024) <https://www.bbc.com/future/article/20241018-ai-art-the-end-of-creativity-or-a-new-movement>.

in key jurisdictions such as the US has contributed to its apparently relatively unregulated development.[19] Lack of transparency from corporate actors as to the origin and use of data input into these models adds to the difficulty of discerning whether their activities can constitute infringement from a practical perspective.[20]

In response to these ongoing developments, the Federal Attorney-General commenced a process of Roundtables in February 2023, resulting in the establishment of a Copyright and AI Reference Group ('CAIRG') in December 2023 to engage in ongoing policy discussions on the implications AI poses to copyright law.[21] The Australian Senate later resolved to establish a Select Committee on Adopting Artificial Intelligence in March 2024 to investigate legal and policy issues arising within the Australian context, that tabled a report in November 2024.[22] This report referred to the 'unprecedented theft' of creative work by multinational technology companies operating in Australia, recommending the use of copyright protected works in training data be 'appropriately licenced', and consultation to consider an 'appropriate mechanism to ensure fair remuneration is paid to creators for commercial AI-generated outputs based on copyrighted material used to train AI systems'.[23] Since then, the Productivity Commission has seemed to advise in the opposite direction, seeking feedback on a proposal to add a new text and data-mining exception to copyright on the basis that it would encourage AI development and use.[24] In this context, a fuller consideration of the complexity in the actual legal position on copyright infringement, in light of the most up-to-date knowledge of the field, is urgently needed.

Moreover, we are now in a position to draw on insights from the arguments offered in the raft of ongoing international litigation, notably in the US and UK, the outcomes of which might conceivably influence domestic Australian courts (despite the notable differences in the copyright framework). Arguments presented in recent cases (in the text-to-image context, notably *Getty Images (US) Inc v Stability AI Inc*,[25] *Kneschke*,[26] and *Andersen v Stability AI Ltd*[27]) offer assistance with new detail on how GenAI interacts with copyright infringement doctrines and potential insights on how courts may approach balancing concerns about inhibiting AI development with copyright's imperative to ensure adequate protection for human-made artistic works.

---

19 Lee, Cooper and Grimmelmann (n 10).

20 Amazon, Meta and Google have declined to answer specific questions on their use of copyrighted material to the Senate Committee: *Senate Report* (n 12) [4.77]–[4.79].

21 'Outcomes Paper on AI and Copyright' (n 12).

22 *Senate Report* (n 12).

23 Ibid. See Recommendations 8, 9 and 10: at 172–3 [1.47]–[1.52].

24 *Productivity Commission Report* (n 7). Just before publication, the Federal Government indicated it did not intend to adopt this proposal, although it is considering other possible copyright reforms: Michelle Rowland, 'Albanese Government to Ensure Australia Is Prepared for Future Copyright Challenges Emerging from AI', *Attorney-General's Portfolio* (Press Release, 26 October 2025) <https://ministers.ag.gov.au/media-centre/albanese-government-ensure-australia-prepared-future-copyright-challenges-emerging-ai-26-10-2025>.

25 *Getty (US) v Stability AI* Complaint (n 3).

26 *Kneschke* (n 5).

27 *Andersen v Stability AI Ltd*, 700 F Supp 3d 853 (ND Cal, 2023) ('*Andersen v Stability AI 2023*').

Alongside the ongoing litigation, the increasing capabilities of GenAI technology have sparked a proliferation of legal scholarship considering its impact on copyright law's fundamental concepts. Some scholarship in this area has previously assumed that the GenAI process is predicated on widespread copyright infringement, largely glossing over the technical steps involved in the infringement analysis to focus on the application of exceptions.[28] New international scholarship provides deeper insights into the stages of the GenAI process.[29] Katherine Lee, A Feder Cooper, and James Grimmelmann demonstrate how a 'supply chain' analysis that breaks down the steps in the process of creating and deploying a GenAI can aid in a more systematic and accurate analysis of the elements of copyright infringement.[30] Considered in this way, liability across each step is by no means clear.[31] In the Australian context, limited consideration has been afforded to the intricate technical processes involved in each stage of the GenAI process, and their relation to the issue of copyright infringement.[32] This article aims to begin addressing this gap.

A specific area of disagreement is whether, and at what stage, a GenAI process might involve 'copying' in a copyright-relevant sense.[33] It is often assumed that 'copies' of works are necessarily made across the stages of the GenAI process, including the inputs, machine learning, and output stages.[34] In the US copyright context, however, some scholars have argued that as the training process involves 'non-expressive' uses of creative works, this cannot, and should not, satisfy the legal requirement for copying, which triggers the reproduction right.[35] Although

28 Matulionyte, 'What Are the Options?' (n 8).

29 Lee, Cooper and Grimmelmann (n 10); Murray (n 14); Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66(2) *Journal of the Copyright Society* 291 <http://dx.doi.org/10.2139/ssrn.3331606> ('The New Legal Landscape').

30 Lee, Cooper and Grimmelmann (n 10).

31 Ibid.

32 Matulionyte, 'What Are the Options?' (n 8); Rita Matulionyte, 'Generative AI and Copyright: Exception, Compensation or Both?' (2023) 134 *Intellectual Property Forum* 33 <https://doi.org/10.2139/ssrn.4652314> ('Exception, Compensation or Both'). Cf Rita Matulionyte, 'Reconceptualising the Reproduction Right in the Age of AI' (Working Paper, 2 December 2024) <https://doi.org/10.2139/ssrn.5041741> ('Reconceptualising the Reproduction Right'). Conceding the infringement questions around AI-output may be more nuanced: Cheryl Foong, 'Immaterial Copying in the Age of Access' (2022) 44(9) *European Intellectual Property Review* 513 <https://doi.org/10.2139/ssrn.4193599>.

33 Murray (n 14); Matulionyte, 'What Are the Options?' (n 8) 420–1; Matthew Sag, 'Copyright Safety for Generative AI' (2023) 61(2) *Houston Law Review* 295 ('Copyright Safety for GenAI'); Sag, 'The New Legal Landscape' (n 29); Lee, Cooper and Grimmelmann (n 10) 324–5; Lemley (n 13); Mark A Lemley and Bryan Casey, 'Fair Learning' (2021) 99(4) *Texas Law Review* 743 <https://doi.org/10.2139/ssrn.3528447>.

34 Matulionyte, 'What Are the Options?' (n 8) 420–1; Matulionyte, 'Exception, Compensation or Both' (n 32) 34–5; Lee, Cooper and Grimmelmann (n 10); Katherine Lee et al, 'AI and Law: The Next Generation: An Explainer Series', *GenLaw* (Web Page, 6 July 2023) <https://genlaw.github.io/explainers/>.

35 Cf *Copyright Act 1968* (Cth) s 31(1)(b)(i) ('*Copyright Act*'). Oren Bracha, 'The Work of Copyright in the Age of Machine Production' [2024] (Fall) *Harvard Journal of Law and Technology* 171, 180. See also Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103(4) *Northwestern University Law Review* 1607, 1626 ('Copyright and Copy-Reliant Technology') and Matthew Sag, 'Orphan Works as Grist for the Data Mill' (2012) 27 *Berkeley Technology Law Journal* 1503 <https://doi.org/10.2139/ssrn.2038889> making the case that non-expressive use should be non-infringing but arguing that this should be operationalised in the US under the umbrella of fair use (an option currently unavailable

the contours of US and Australian copyright law are not exactly the same, we argue that the idea of non-expressive use might be a useful conceptual tool. It could allow courts to interpret the ambit of the reproduction right in a way that accounts for its normative role within copyright law as a key means for designating the particular interactions with creative work that should be regulated to support the human creative ecosystem that copyright is intended to underpin and encourage.[36]

With GenAI's rise in modern society, current policy discussions in Australia and globally highlight the difficulty of achieving balance between encouraging the development of text-to-image AI models and adequate protection of artists' rights.[37] Were Australian courts to favour a more liberal approach to the use of copyright works, this may encourage greater development in AI technology domestically.[38] In this way, Australia may be better placed to offer a viable alternative to the major global players that dominate in the race to develop and disseminate this new technology. Considering the anticipated growth of the AI industry, some have argued that restricting the use of copyrighted content in machine learning processes disadvantages Australia economically.[39]

Beyond the economic effect, it may also hinder efforts to create a 'sovereign' AI capacity that reflects our needs and serves security interests.[40] Most significantly, restricting non-expressive access to training data may contribute to embedding harmful biases in models that risk broader harms to Australian society in some cases.[41] Ensuring that these powerful technologies are not predicated on widespread bias is a significant public interest issue.[42] Of course, all of this must be balanced with the potentially significant disruption to creators and the creative industries caused by flooding the market with low-cost substitutes seen to free-ride on their work. The stakes are high as this potentially impacts not only the shape of the creative labour market, but also the content and contours of our cultural sphere.

The regulation of GenAI is of significant economic, cultural, and social importance. But, it is also worth remembering that it is not the first technological advancement to challenge the tenets of copyright law, nor will it be the last.[43] To retain a strong and relevant copyright law system, Australian courts should maintain

---

in the Australian context). Cf Robert Brauneis, 'Copyright and the Training of Human Authors and Generative Machines' (2024) 48(1) *Columbia Journal of Law and the Arts* 1 <https://doi.org/10.52214/jla.v48i1.13529> questioning the usefulness of the concept of 'non-expressive' use.

36 The sense in treating 'non-expressive' uses as non-infringing was recognised by the Australian Law Reform Commission ('ALRC'): Australian Law Reform Commission, *Copyright and the Digital Economy. Final Report* (ALRC Report No 122, November 2013) 260–4 ('*Copyright and the Digital Economy Report*'). This was discussed in the context of proposed reform to copyright exceptions.

37 'Outcomes Paper on AI and Copyright' (n 12) 3.

38 Matulionyte, 'What Are the Options?' (n 8).

39 Ibid 418.

40 The current government includes AI sovereign capability amongst its priorities: see Alex Antic, 'Owning the Algorithm: Australia's Path to AI Sovereignty', *InnovationAus.com* (online, 26 May 2025) <https://www.innovationaus.com/owning-the-algorithm-australias-path-to-ai-sovereignty/>.

41 Lee, Cooper and Grimmelmann (n 10) 292.

42 Noting the problem of bias: see Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93 *Washington Law Review* 579.

43 Daniel Gervais et al, 'The Heart of the Matter: Copyright, AI Training, and LLMs' (2024) 71(3) *Journal of the Copyright Society* 482, 516 <https://doi.org/10.2139/ssrn.4963711>.

a *principled* approach to considerations of copyright infringement that upholds copyright's core purpose to encourage creativity and regulate the communication of human expression. We will argue that a purely economic approach focused on marketplace incentives cannot capture the full picture. Before this, we turn to the doctrinal analysis.

III GENAI AND THE COPYRIGHT INFRINGEMENT ANALYSIS

A systematic approach is needed to clearly address the question of whether liability for copyright infringement can be attributed to the human actors at each of the stages in the GenAI process. Lee, Cooper, and Grimmelmann carefully break down what they term the GenAI 'supply chain' into eight stages: the creation of expressive works, data creation, dataset collection and curation, model (pre-) training, model fine-tuning, model release and system deployment, generation, and model alignment.[44] They explain how each stage is interconnected and their article is a rich resource, incorporating new technical detail about the characteristics of each stage and focusing on identifying each choice made by a company or user that might have copyright relevance. For simplicity, we gather these together in four key stages, selected because they require separate analysis from a copyright infringement point of view. These are:

(a) Creating training datasets;
(b) Machine learning;
(c) Foundational models; and
(d) Output (prompting AI models and generating images).[45]

This Part will consider the Australian test for copyright infringement and its application to the above four stages of the text-to-image GenAI supply chain. After first delineating the parties involved, each section will examine whether the activities undertaken by various actors satisfy the elements of the copyright infringement test: an infringing act; the taking of the whole or a substantial part of a protected work; and a sufficient causal connection to substantiate a finding of infringement.[46] Where the parties involved could be considered to infringe upon the copyright subsisting in protected works, for completeness, the application of relevant exceptions will be briefly considered, specifically the fair dealing and temporary reproduction exceptions.[47]

This supply chain approach is particularly useful given the reality that, in many cases, aspects of the creation, training, and deployment of AI models occur in different places and are undertaken by different actors. Where each step occurs is likely to depend upon considerations such as the availability of

44 Lee, Cooper and Grimmelmann (n 10) 285–8. There will also be an element of model planning, design and alignment involved.
45 See ibid 286, figure 6.
46 Kathy Bowrey et al, *Australian Intellectual Property: Commentary, Law and Practice* (Oxford University Press, 3rd ed, 2021) ch 6; *Copyright Act* (n 35) ss 14, 31(1)(b)(i), (iii).
47 *Copyright Act* (n 35) ss 40–3B.

computing infrastructure, labour costs, and the presence of a favourable regulatory environment. Choices made by actors at one point in the process might affect the liability of actors at other stages. Currently, many potentially significant infringing reproductions (such as creating training data and training models) occur outside Australia. Nonetheless, local fine-tuning of AI models, their deployment, and the generation of outputs, might constitute infringing acts in Australia.[48] Further, the degree to which Australia offers a favourable location for all the stages of the AI supply chain is a material consideration in current discussions about developing sovereign Australian AI capability.[49]

A Creating Training Datasets

At what is often described as the 'input' stage of the GenAI process, two key activities are undertaken: training dataset creation and curation, then the use of training data in machine learning for model pre-training.[50] These steps will be discussed in this section and the next. Usually, two main parties might be involved at these stages for the purposes of attributing liability: AI model developers themselves and potentially also a third-party dataset creator.[51] There is significant risk that the data used for machine learning reproduces copyright protected works, potentially infringing the exclusive rights of copyright holders,[52] such as their right to reproduce the work in material form[53] and their right to communicate their work to the public.[54] This section focuses on the processes of dataset creation and curation. The next section will consider the use of datasets in machine learning processes.[55]

The process of training dataset creation involves compiling likely billions, or even trillions, of pre-existing image files and their alt-text into datasets that are then cleaned and processed, ready for use in AI model training.[56] Alt-text is a short description associated with an image on a webpage that conveys its meaning and

48 Our analysis of the machine learning stage applies to activities such as fine-tuning of artificial intelligence models.

49 Antic (n 40). On the importance of considering copyright regulatory settings and 'copyright economics' as part of ensuring sovereignty in the AI supply chain and homegrown models in the EU context: Bertin Martens, 'The EU's False Sense of Isolationism in AI and Copyright', *Kluwer Copyright Blog* (Blog Post, 29 May 2025) <https://copyrightblog.kluweriplaw.com/2025/05/29/the-eus-false-sense-of-isolationism-in-ai-and-copyright/>.

50 Lee, Cooper and Grimmelmann (n 10) 257–8.

51 Ibid 291.

52 Matulionyte, 'Exception, Compensation or Both' (n 32) 3.

53 *Copyright Act* (n 35) s 31(1)(b)(i).

54 Ibid s 31(1)(b)(iii).

55 Here the issue is reproduction of the work in material form, rather than its communication to the public: see below Part III(B).

56 Kimberlee Weatherall, 'Technology and the Law: Managing Intellectual Property and Copyright in a World of AI' (Seminar, Australian Council of Learned Academies Parliamentary Library Seminars, 21 February 2024) <https://www.aph.gov.au/About_Parliament/Parliamentary_departments/Parliamentary_Library/Research/Lectures?searchTerms=weatherall&selectedVideo=%7BA73878A3-897F-4BC5-8AA2-3EDCA4A320BA%7D>.

context.[57] A portion of the data in training datasets may exist in the public domain and thus be available for unrestricted use.[58] But, it is generally acknowledged that the majority of works used in training AI models – particularly in the development of large, foundational models – are protected by copyright and have been scraped from the internet without express authorisation.[59] The dataset creation and curation process primarily implicates the reproduction right.[60] To prove that an infringing act occurred during the inputs stage in Australia, it is necessary to establish that there is a reproduction of the artistic work *in a material form*.[61]

The Australian understanding of the reproduction right has evolved in the wake of the digital era. Section 21(1A) of the *Copyright Act 1968* (Cth) ('*Copyright Act*') elaborates the definition of 'reproduction' by providing that 'a work is taken to have been reproduced if it is converted into … a digital or other electronic machine-readable form'.[62] Further, the definition of 'material form' explicitly provides for 'any form (whether visible or not) of storage of the work', which includes digital copies in a computer's memory.[63] In this way, the *Copyright Act* has been understood to codify an expansive view of digital copies of artistic works that includes both temporary and long-term copies stored on the devices of dataset creators.[64]

At the dataset creation stage, a model developer usually downloads 'permanent files' of the images. These will be cleaned and processed as part of the larger dataset for use in model training.[65] As Matthew Sag explains, whilst it is technically possible for GenAI models to be trained on datasets that are based in cloud-storage services, it is more resource-efficient for larger companies to store image information files on local devices.[66] When this happens, the party that undertakes the activity of creating 'digital copies' within their local storage will likely infringe

57 'Everything You Need to Know to Write Effective Alt Text', *Microsoft Support* (Web Page) <https://support.microsoft.com/en-au/office/everything-you-need-to-know-to-write-effective-alt-text-df98f884-ca3d-456c-807b-1a1fa82f5dc2>.

58 Sag, 'Copyright Safety for Gen AI' (n 33) 315.

59 Ibid; Ed Newton-Rex, 'How AI Models Steal Creative Work: And What to Do About It' (Speech, TEDAI San Francisco, 22 October 2024) <https://www.youtube.com/watch?v=U9d0p96N1iw>. This includes activities such as web scraping, web crawling and screen scraping: see Organisation for Economic Co-operation and Development, 'Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data' (Artificial Intelligence Paper No 33, 9 February 2025) <https://www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_d5241a23-en.html> ('OECD Report').

60 Sag, 'Copyright Safety for Gen AI' (n 33) 295; *Copyright Act* (n 35) s 31(1)(b)(i).

61 *Copyright Act* (n 35) s 31(1)(b)(i).

62 Ibid s 21(1A).

63 Ibid s 10(1). This definition is specified to apply regardless of whether or not the work can be reproduced from this form of storage.

64 Bowrey et al (n 46) 210. This conclusion follows from the Explanatory Memorandum to the *US Free Trade Agreement Implementation Act 2004* (Cth) explaining the change to the definition of 'material form' as implementing article 17.4.1 of the Australia-US Free Trade Agreement: at Explanatory Memorandum, US Free Trade Agreement Implementation Bill 2004 (Cth) 149 [668].

65 Matthew Sag, 'Fairness and Fair Use in Generative AI' (Peter A Jaszi Distinguished Lecture on Intellectual Property, Program on Information Justice and Intellectual Property, 28 September 2023) <https://www.youtube.com/watch?v=805O3kGj4_Y> ('Fairness and Fair Use').

66 Ibid.

copyright.[67] By downloading files to their local servers, model developers and training dataset compilers necessarily reproduce the works in a 'material form' that can be compared to a digital copy on a computer's memory, satisfying section 10(1) of the *Copyright Act*.[68] As the whole image is downloaded, there is no need to investigate the 'substantial part' element in proving infringement.[69] In the absence of express authorisation of the copyright owner, the creation of these digital copies would likely constitute infringement of the exclusive right to reproduction.[70]

In some situations, the right to communicate works to the public will also be implicated.[71] But this may not always be straightforward. This is aptly illustrated with reference to the Large-scale Artificial Intelligence Open Network dataset ('LAION-5B'), which is a collation of web-scraped images and their text-data, shared via a website.[72] Containing over 5 billion text-image pairs, subsets of LAION-5B have been used to develop leading text-to-image GenAI models, including Midjourney and Stable Diffusion.[73] As one of the relatively few publicly available, large-scale datasets, LAION-5B has been commented on and its use has been subject to litigation.[74]

Making images available online may constitute the communication to the public of those works.[75] 'Communicate' is defined in section 10(1) as '[making] available online or electronically [transmitting] (whether over a path, or a combination of paths, provided by a material substance or otherwise)'. This definition was drafted to ensure its technological neutrality and is considered to encompass both 'push' electronic communications (such as broadcasting) and 'pull' communications (such as making a work available on an internet platform so that it might be accessed on demand).[76] A relevant communication is deemed to have been undertaken by 'the person responsible for determining the content of the communication'.[77] Where a platform provides access to content, whether the person responsible for the communication is the person providing the works to the platform, the platform itself, those accessing works via the platform, or some combination, is unclear.[78]

---

67 Ibid.
68 *Copyright Act* (n 35) s 10(1).
69 Ibid s 14(1).
70 Ibid ss 31(1)(b)(i), 36(1).
71 Ibid s 31(1)(b)(iii).
72 Christoph Schuhmann, et al, 'LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-text Models' (Conference Paper, Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks, 16 October 2022) <https://doi.org/10.48550/arXiv.2210.08402>.
73 Ibid.
74 Andres Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73(2) *GRUR International: Journal of European International IP Law* 111 <https://doi.org/10.1093/grurint/ikad140>; Brauneis (n 35) 32–5; *Kneschke* (n 5).
75 *Copyright Act* (n 35) s 31(b)(iii).
76 David Brennan, *Copyright Law* (Federation Press, 2021) 173 quoting the Explanatory Memorandum, Copyright Amendment (Digital Agenda) Bill 1999, 24.
77 *Copyright Act* (n 35) s 22(6).
78 Brennan (n 76) 179–80.

In the case of LAION-5B, the position is complex. In German litigation, it was argued that this dataset does not make images themselves available.[79] This is because it only contains alt-text and hyperlinks to where the images could be found when the dataset was created.[80] In Australia, the Federal Court has determined that merely providing a hyperlink to third-party hosted content will not be sufficient to count as the *communication* of that content to the public.[81] Further, copying or communicating merely the image-text data present on the online database likely falls short of a 'substantial part' of a work.[82] This data is arguably not even part of the work itself.[83] Stored image-text data is metadata associated with the image – not an aspect of the image itself – and, as such, it cannot be considered a substantial part of a protected artistic work.[84] The image-text data collected for each image is confined to the URLs (website locations) of each image and associated alt-text (metadata that can be likened to rich captions describing each image).[85]

Whilst databases of image-text data themselves can be analogised to large catalogues of the images on the internet (describing but not containing the images), consideration still needs to be undertaken of the manner in which data is collected using these databases and whether infringing copies are necessarily made.[86] Where a model developer locates and downloads image files using the shared URLs, an infringing copy would be created – even if it is not 'communicated' to the public by making this database publicly available.[87] And, the finding that matters most for this stage of our analysis is that the process of actually making a dataset like LAION-5B likely infringes the reproduction right of copyright holders.[88] In German litigation, Large-scale Artificial Intelligence Open Network ('LAION') (the not-for-profit organisation behind LAION-5B) admitted to making copies of images for the purpose of extracting information, though these were subsequently deleted.[89] The result in that case was that these activities were permissible as they

79 *Kneschke* (n 5).

80 Ibid.

81 *Universal Music Australia Pty Ltd v Cooper* (2005) 150 FCR 1, 18 (Tamberlin J). A hyperlink merely 'facilitates the easier location' of a protected work, it does not transmit or make it available: at [65]. Similarly, accessing such a link is not an act of communication: see *Copyright Act* (n 35) s 22(6A) and discussion in Brennan (n 76) 178–9.

82 *Copyright Act* (n 35) s 14(1). Alt-text captions are too insubstantial and lack the originality required to be protected as literary works themselves.

83 In *Interlego AG v Tyco Industries Inc* [1989] AC 217, the Privy Council held that annotations on a drawing (indicating the dimensions of a Lego block) were not part of the artistic work protected by copyright.

84 Sag, 'Copyright Safety for Gen AI' (n 33) 308.

85 Ibid 303 citing Romain Beaumont, 'LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets', *LAION* (Blog Post, 31 March 2022) <https://laion.ai/blog/laion-5b/>; Murray (n 14); Lee, Cooper and Grimmelmann (n 10) 291 n 138; Lee et al (n 33) 9. Guadamuz (n 74) has argued that weighing against a finding of communication to the public may also be the fact that entries to the database are difficult to access (in the case of LAION-5B it includes over 5 billion links) – it takes significant effort to download the entire database, find a specific link and access it.

86 Weatherall (n 56).

87 Murray (n 14); Lee, Cooper and Grimmelmann (n 10) 283–4.

88 *Copyright Act* (n 35) ss 10(1), 31(1)(b)(i).

89 *Kneschke* (n 5) 6.

fell within the scope of EU data mining exceptions considering the 'scientific' purpose of the dataset[90] (an exception which presently has no Australian equivalent – a point we return to in Part IV).

Another dimension that should be considered is that the dataset creation and curation process often occurs internationally, as is the case for LAION-5B. The potentially global nature of elements of the GenAI supply chain generates jurisdictional issues as suggested above. Currently, much of the copying required for the training process for many GenAI models is undertaken on servers overseas, meaning that any potentially infringing conduct at this stage will rarely fall under the jurisdiction of Australian copyright law.[91] Understandably, AI model developers and companies will prioritise jurisdictions with more favourable legal frameworks such as Singapore, Japan, or the US.[92]

The risk that this will impede the growth of Australia's technology industry has been noted in official reports.[93] Practically speaking, the training datasets used for image-generating AI available in Australia will be compiled overseas in jurisdictions with broader copyright exceptions to limit liability risks. Although primary infringement activities are undertaken overseas, text-to-image GenAI models may then be 'imported' into Australia. This does raise questions about liability under Australian parallel importation provisions.[94] Where it can be established that a copy of the protected work was made without a licence from the owner in the country where the training dataset was made, it may be open to a court to find that the trained text-to-image GenAI imported into Australia infringes copyright.[95] This is facilitated by the application of a knowledge standard that is defined by reference to Australian copyright law standards for infringement (rather

90 *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC* [2019] OJ L 130/92, art 3, implemented in *Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes* [Law to Adapt Copyright Law to the Requirements of the Digital Single Market] (Germany) 13 May 2021, BGBl, 2021, 1204.

91 Weatherall (n 56); 'OECD Report' (n 59) 25.

92 Singapore and Japan have recently amended their laws to include AI-friendly exceptions: *Copyright Amendment Act 2021* (Singapore) introduced a new section 244 exception for the use of copyright protected works for 'computational data analysis', and Japan amended its copyright law to include a new article 30–4 which applies to non-enjoyment uses of a work, discussed further in Part IV below. The volume of ongoing litigation in the US tests the limits of the fair use exemption which many regard as potentially extending to uses of copyright protected works in AI training. For a detailed discussion and jurisdictional comparison of potentially relevant exceptions: see Matthew Sag and Peter K Yu, 'The Globalization of Copyright Exceptions for AI Training' (2025) 74(5) *Emory Law Journal* 1163 <http://dx.doi.org/10.2139/ssrn.4976393>.

93 *Senate Report* (n 12); Weatherall (n 56).

94 *Copyright Act* (n 35) ss 37–8. Jani McCutcheon, 'Are Generative AI Models Infringing Imported 'Articles' Under Copyright Law?' (2025) *Sydney Law Review* (forthcoming) <https://ssrn.com/abstract=4811683>, provides a useful analysis of the application of these provisions to AI systems, which were not designed to apply to digital products accessed online. See also Mattias Rättzén, 'Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI' (2024) 26(1) *Columbia Science and Technology Law Review* 175, 228 <https://doi.org/10.52214/stlr.v26i1.13338>.

95 McCutcheon (n 95).

than overseas standards).[96] Still, establishing liability may be challenging due to the evidentiary difficulty in actually proving that an infringing copy was made and used in the training process of the text-to-image GenAI model in question. Furthermore, these models tend to be accessed by Australians online rather than being *physically* imported (arguably the activity parallel importation provisions are designed to capture).[97] As Part III(C) below will explain, arguably there is no final 'copy' of works in the training data stored in a final AI model itself. This strengthens the argument that there is no parallel importation infringement liability.

B Machine Learning

Another key process in the development of a text-to-image GenAI model is machine learning, which refers to the processes by which a model developer guides an AI system to build its capacity to perform desired tasks.[98] These computational processes have been claimed to involve countless iterations of copying of the image and alt-text files in the training dataset.[99] Once the dataset with images and alt-text data or captions is compiled, machine learning generally involves 'feeding' the model a pre-processed training dataset, triggering investigation into infringement of the reproduction right.[100] As there is no 'electronic transmission' or 'making available' to the public in machine learning, the right to communicate the work to the public is not engaged.[101] This section will focus on diffusion-based machine learning, as diffusion models are at the core of some of the most popular text-to-image GenAI models today.[102]

Distinctions between the data preparation and training processes for various types of machine learning demonstrate the necessity of avoiding labelling and examining GenAI infringement in a generalised manner. Each instance needs to be considered individually as specific technical processes raise different issues.[103] Consequently, the results of a copyright infringement analysis may vary from

96 *Copyright Act* (n 35) s 37(1): 'if the importer knew, or ought reasonably to have known, that the making of the article would, if the article had been made in Australia by the importer, have constituted an infringement of the copyright'. See also section 38 in respect of the subsequent sale of such articles.

97 McCutcheon (n 94) discusses the potential meaning of 'importation' in this context, providing a nuanced view of whether it can be extended to these sorts of circumstances (considering it desirable to extend the concept in these circumstances).

98 BJ Ard, 'Copyright's Latent Space: Generative AI and the Limits of Fair Use' (2025) 110(3) *Cornell Law Review* 509, 522–4.

99 As stated above, an image's alt-text refers to a caption or information to assist with classifying the image. Benjamin L W Sobel, 'Artificial Intelligence's Fair Use Crisis' (2017) 41(1) *Columbia Journal of Law and the Arts* 45; 'Outcomes Paper on AI and Copyright' (n 12); Matulionyte, 'What Are the Options?' (n 8) 419.

100 *Copyright Act* (n 35) ss 10(1), 31(1)(b)(i).

101 Ibid ss 10(1), 31(1)(b)(iii).

102 For example, DALL-E 3, Stable Diffusion, Midjourney, Google's Imagen.

103 A Feder Cooper and James Grimmelmann, 'The Files Are in the Computer: On Copyright, Memorization, and Generative AI' (2025) 100(1) *Chicago Kent Law Review* 141, 156 <https://doi.org/10.48550/arXiv.2404.12590>.

model to model.[104] Training a diffusion-based machine learning model[105] generally involves three steps:

1. Pre-processing and encoding the image and text files into lower-dimensional formats (smaller and containing fewer variables);[106]
2. Forward-diffusion by adding 'noise' to encoded images; and
3. Reverse diffusion on the 'noisy' image to gradually generate a novel output.[107]

The first step in the machine learning process involves pre-processing the image files and their corresponding alt-text captions to ensure appropriate formatting for model training.[108] At this stage, the model developer likely has a copy of the training dataset in some form to enable data pre-processing and encoding.[109] Auto-encoders can be used to compress original image files into lower-dimensional representations, containing only the most important information from the image for storage in the model's latent space.[110] A model's 'latent space' is a spatial representation of the patterns and correlations a model learns from its training data. The original images are essentially extractible from these 'latent space vectors' through the use of a 'decoder'.[111] Similar to the compression undertaken by a .zip file, a decoder aims to reconstruct the original image as closely as possible.[112]

In models such as Stable Diffusion, however, this dimensionality reduction is conducted by specific types of auto-encoders called variational auto-encoders ('VAEs').[113] As opposed to standard auto-encoders, VAEs are used to compress images into representations that extract the statistical properties of the original images.[114] By extracting the image's key features, these latent space vectors allow for greater computational efficiency during machine learning.[115] Latent space vectors encoded by VAEs have been explained as 'perceptually equivalent to the

104 On the range of approaches: see Guadamuz (n 74).

105 Rombach et al, 'High Resolution Image Synthesis with Latent Diffusion Models' (Conference Paper, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27 September 2022) <https://doi.org/10.1109/CVPR52688.2022.01042>. These lower-dimensional representations are referred to as 'latent space vectors'.

106 William Peebles and Saining Xie, 'Scalable Diffusion Models with Transformers' (Conference Paper, ICCV Conference, 2 March 2023) 4 <https://arxiv.org/abs/2212.09748>.

107 Rättzén (n 94) 184–91.

108 'OECD Report' (n 59).

109 Guadamuz (n 74) 115.

110 Rombach et al (n 105) 10,685; Rättzén (n 94) 189; Dave Bergmann and Cole Stryker, 'What is an Autoencoder?', *IBM* (Web Page, 23 November 2023) <https://www.ibm.com/think/topics/autoencoder>.

111 Bergmann and Stryker (n 110).

112 Nevertheless, these versions are still 'lossy' in the sense that they reduce file size by permanently removing some of the original data.

113 See generally, Rombach et al (n 105) 10,685–6.

114 Ard (n 98) 531–4.

115 Bergmann and Stryker (n 110). Image files are 'passed' through a VAE to compress standard image files 256x256 pixels in size into lower-dimensional representations in the latent space of the model. This process of 'compressing' images into lower-dimensional, numerical representations effectively reduces the size of the training dataset, enabling processing efficiencies when working with very large datasets. The text captions are also encoded into the model using the OpenCLIP-ViT/H text-encoder. Images encoded through VAEs provide a more stable latent space for diffusion and enable greater generative capabilities in comparison to traditional auto-encoders.

image space'.[116] Arguably, though, a level of transformation of the training dataset occurs, raising questions as to whether they 'reproduce' a 'substantial part' of the original work. This is because VAEs learn compressed representations of their training data as probability distributions, which allow the important features from the images input to be encoded in an approximate – rather than in a deterministic or directly representational – way in contrast to standard auto-encoders.[117]

During the second step, a Stable Diffusion model takes these latent space vectors and adds 'noise' to gradually degrade them in a process of hundreds to thousands of steps, namely a diffusion process in its latent space.[118] Essentially, by exposing the model to how an image degrades over time, the model learns the features of images linked with particular text. Then, in the third step, the model learns to reconstruct the original data or generate similar iterations by reversing this process of noise addition. Significantly, the latent space vectors used to enable machine learning are only stored for the duration of model training and are not retained permanently once training is completed.[119] The only information retained from the original images is the learned statistical correlations stored in the model's weights, enabling generation of new images.[120]

The encoding of latent space vectors in the diffusion-based machine learning process may be seen to create temporary copies for the purposes of machine learning. Despite only existing for a transitory duration as an intermediate step in the machine learning process, their use in the diffusion processes may satisfy an infringing act of reproduction in a material form, which is usually understood to include temporary copies.[121] Findings on this issue may depend on how deeply a court will engage with the technical process and especially its view of the level of transformation that has occurred (ie, is there sufficient objective similarity for copying or reproduction to be said to have occurred). Does this process count as the conversion of the work 'into … a digital or other electronic machine-readable form' (included in the definition of reproduction in section 21(1A))?

The expanded definition of 'material form' adopted to implement the Australia-US Free Trade Agreement (that is, expressed to apply regardless of whether a work is reproducible from electronic storage) may seem to indicate an extraordinarily broad scope to the reproduction right that could extend to machine learning processes.[122] Yet, in expanding the meaning of 'material form' this amendment did not alter the central importance for there to be a 'reproduction' in the first place.

---

116 Rombach et al (n 105) 10,866. See also Guadamuz (n 74) 113–4.

117 Rombach et al (n 105).

118 Lee, Cooper, Grimmelmann (n 10) 281–3; Dave Bergmann, 'What is Latent Space?', *IBM* (Web Page, 28 January 2025) <https://www.ibm.com/think/topics/latent-space>.

119 Guadamuz (n 74) 115–7; Rättzén (n 94) 190–1; Matulionyte, 'Reconceptualising the Reproduction Right' (n 32) 15–17.

120 Guadamuz (n 74) 114.

121 *Copyright Act* (n 35) ss 31(1)(a)(i), (b)(i), read in light of the section 10(1) definition of 'material form' and the section 43A, 43B, 111A, 111B exceptions for temporary copies in certain circumstances (implying temporary copies would otherwise fall within the scope of the reproduction right). See also Sobel (n 99) 62–3.

122 Discussing this view and arguing to the contrary: see Brennan (n 76) 152.

Section 21(1A) is silent on reproducibility. In determining whether a reproduction has occurred, it remains possible, at least theoretically, for courts to take into consideration whether steps can be taken to reproduce or extract the work from the alleged copy. Perhaps the novelty of GenAI technology could support such an approach, if backed by technical evidence.[123]

In the process of machine learning, the 'copies' used and contained in the training datasets will only ever be 'read' or consumed by a machine for statistical analysis and not appreciated by a human audience.[124] Machine learning itself does not take from the work the innate value that copyright arguably protects, its unique expressive value that enables appreciation by an audience.[125] The mathematical information a model takes from its training data does not form a critical part of the *communicative* aspect of a creative work. As Mark Lemley and Bryan Casey express it, the value the AI model takes from the protected work stems from the part of the work that copyright law has decided is in the public domain (information, ideas).[126] Sag provides a strong argument, albeit in the US context, that copyright should only protect uses of works that are expressive, that is, uses that are 'intended to enable human enjoyment, appreciation, or comprehension of the copied expression as an expression'.[127]

An argument might conceivably be made that 'reproduction' should be understood in light of its role in the context of the *Copyright Act*. In a dispute on this question, courts may be forced to probe the normative dimensions of

123 This line of reasoning successfully negated a finding of infringement based on an earlier definition of 'material form': see eg *Stevens v Kabushiki Kaisha Sony Computer Entertainment* (2005) 224 CLR 193 ('*Stevens v Kabushiki*'). Such an approach is excluded by the current definition of 'material form' as amended to implement the Australia-US Free Trade Agreement. One might argue that this change does not preclude understandings of 'reproduction' that take into account reproducibility from the alleged copy. The Australia-US Free Trade Agreement required Australia to extend protection to 'temporary' reproductions, including temporary storage in material form. Two observations might be made. First, it is not clear that this extends to the storage of statistical representations of attributes relating to a work. Second, even if this was intended, it may be that this provision of the Agreement has been incompletely implemented. The legislature chose to amend the definition of 'material form', leaving the definition of 'reproduction' unmodified when it comes to reproducibility. This notably *indirect* approach may cause one to wonder if the effect is to preserve some wriggle room, given the centrality of the concept of reproduction to copyright law.

124 Sag, 'Copyright Safety for GenAI' (n 33) 307–10.

125 Carys J Craig, *Copyright, Communication and Culture: Towards a Relational Theory of Copyright Law* (Edward Elgar Publishing, 2011) ch 4 <https://doi.org/10.4337/9780857933522.00002> ('*Copyright, Communication and Culture*); Sag, 'Copyright Safety for GenAI' (n 33); Lemley and Casey (n 33) 767–8. It is not use of the work '*as a work*': Alain Strowel, 'Reconstructing the Reproduction and Communication to the Public Rights: How to Align Copyright with Its Fundamentals' in P Bernt Hugenholtz (ed), *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Wolters Kluwer, 2018) 203, 207. Interestingly, in *Bartz v Anthropic* (ND Cal, No C 24-05417 WHA, 23 June 2025) slip op 8 ('*Bartz v Anthropic*'), District Judge Alsup compared Anthropic's use of books to a 'reader aspiring to be a writer' in a seeming analogy to human learning. Arguing for the need to resist the temptation to anthropomorphise AI: see Daniela Simone, '"Pay No Attention to that Man Behind the Curtain!": Copyright, Authorship and Artificial Intelligence' (2023) 33(3) *Australian Intellectual Property Law Journal* 120.

126 Lemley and Casey (n 33) 785.

127 Sag, 'The New Legal Landscape' (n 29); Sag, 'Copyright and Copy-Reliant Technology' (n 35). See also Foong (n 32) 19 and Bracha (n 35).

copyright law in new ways. Here, Australian courts are not hamstrung like the Court of Justice of the European Union, for example, by the statutory injunction to give 'reproduction' a broad interpretation.[128] In responding to the challenge that GenAI poses, Carys Craig argues that copyright law should reflect the value society places on the *human* communication and cultural exchange facilitated through the copyright system.[129] She stresses that if copyright is to maintain its purpose as a justifiable limitation on expressive activities, it should not obstruct relations of communication but rather encourage them.[130] Copyright should leave space for non-expressive uses of creative works.[131] In the context of text-to-image AI models, Craig argues that the machines involved do not engage in a relation of communication comparable to a human creator and human audience.[132]

Craig's account is by no means the only to integrate the cultural and communicative dimension of copyright in understanding its scope.[133] The point might be put slightly differently. Where no exchange of meaning takes place, as occurs in machine learning, arguably GenAI does not usurp the communicative value of the protected work *as a* copyright work. In a similar vein, Maurizio Borghi and Stavroula Karapapa have argued that 'de-intellectualized' uses of works should not be impeded by copyright law. Writing about mass digitisation, at a time before GenAI exploded onto the public scene, they explain how automated processes cut the 'intellectual tie' between author and work.[134] In their view, automated processing entails uses of works *qua* data containers in a way that is new to copyright and does

128 The definition of 'material form' is expansive but this should not in itself imply a broader scope of 'reproduction', as discussed below in Part III(C). In the EU, infringement is often considered a foregone conclusion when it comes to AI. Cases like *Infopaq International A/S v Danske Dagblades Forening* (Court of Justice (Fourth Chamber), C-5/08, 16 July 2009) have applied an expansive interpretation of the reproduction right in light of the objective to secure a 'high level' of protection across the harmonised regime, as indicated by the *Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society* [2001] OJ L 167/10, [4]. See Eleonora Rosati, 'Copyright Reformed: The Narrative of Flexibility and Its Pitfalls in Policy and Legislative Initiatives (2011–2021)' (2023) 31(1) *Asia Pacific Law Review* 33 <https://doi.org/10.1080/10192557.2022.2117482>; Eleonora Rosati, 'Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law' [2024] *European Journal of Risk Regulation* 1 <https://doi.org/10.1017/err.2024.72>. The EU approach to reproduction has been criticised: see Séverine Dusollier, 'Realigning Economic Rights with Exploitation of Works: The Control of Authors Over the Circulation of Works in the Public Sphere' in P Bernt Hugenholtz (ed) *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Wolters Kluwer, 2018) 163, 163, 167 <https://ssrn.com/abstract=3544229>.

129 Carys J Craig, 'The AI-Copyright Challenge: Tech-Neutrality, Authorship, and the Public Interest' in Ryan Abbott (ed), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar Publishing, 2022) 134 ('AI-Copyright Challenge').

130 Craig, *Copyright, Communication and Culture* (n 123) 4, 150.

131 Craig, 'AI-Copyright Challenge' (n 129) 136–7; Sag, 'Copyright Safety for GenAI' (n 33) 6–7.

132 Craig, 'AI-Copyright Challenge' (n 129) 145–6.

133 On some such approaches and their philosophical lineage: see Dusollier (n 128) 164–6. For a particularly influential account: see Abraham Drassinower, *What's Wrong with Copying?* (Harvard University Press, 2015) <https://doi.org/10.4159/9780674286566>.

134 Maurizio Borghi and Stavroula Karapapa, *Copyright and Mass Digitization* (Oxford University Press, 2013) ch 7 <https://doi.org/10.1093/acprof:oso/9780199664559.001.0001>.

not fit well within its categories.[135] If copyright protects the expressive nature of works, then some say it stands to reason that the legal meaning of 'reproduction' should not extend to such technical uses of a work.[136]

Thomas Margoni and Martin Kretschmer explain the appropriateness and desirability of such an approach, writing in the EU context, explicitly on the reproduction right. They argue that because copyright only protects *expression*, a broad view of reproduction that encompasses computational uses of works such as machine learning 'frustrates and renders ineffective some of the most important fundamental copyright principles, such as the idea/fact/expression dichotomies'.[137] In the US context, Oren Bracha takes the argument further by cautioning against a 'mechanical' application of the reproduction right without regard to how it serves its underlying purpose in a new context.[138] He reminds that copyright was not meant to exclude or control *all* valuable uses of a work; instead, it is limited by domain and scope.

This vision of copyright aligns with the classic statements of the High Court of Australia in *Victoria Park Racing and Recreation Grounds Co Ltd v Taylor* resisting the extension of intellectual property protection to 'all the intangible elements of value'.[139] And, more recently in *IceTV Pty Ltd v Nine Network Australia Pty Ltd*, that copyright does not protect effort per se, but rather embodies a specific 'social contract' between authors and the reading public.[140] It is clear that copyright should not be reduced to a protection against unfair competition. Part of what takes copyright beyond unfair competition, we would argue, is its concern for the communicative dimension of expression.[141] Machine learning does not, and is not intended to, reproduce the communicative aspects of protected expression in a copyright-relevant sense.[142]

---

135 Ibid. Considering that this sort of use does not 'fit well in the established understanding of what a "use" of a copyright work is': at 141.

136 Borghi and Karapapa (n 134) ch 3 considers automated processing to provide a fundamental challenge to copyright's notion of reproduction, such that 'notions of the "copy" and "copying" have lost their capacity to orient and organize the analysis of infringement': at 51. Sag, 'Copyright Safety for GenAI' (n 33) 304. Brauneis (n 35) criticises the use of 'expressive' in this context on the basis that commentators use it in several different senses that, in his view, are liable to confuse.

137 Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71(8) *GRUR International* 685, 698 <https://doi.org/10.1093/grurint/ikac054>.

138 Bracha (n 35) 7. Dusollier argues that technical applications of the reproduction right in the EU context have led it astray from its origin and purpose to prevent exploitation of copyright works (whilst leaving their 'reception or reading' free): Dusollier (n 127) 167.

139 (1937) 58 CLR 479, 509 (Dixon J) ('*Victoria Park Racing*').

140 (2009) 239 CLR 458, [25], [49]–[52] (French CJ, Crennan and Kiefel JJ) ('*IceTV*'). See also *JR Consulting & Drafting Pty Ltd v Cummings* (2016) 329 ALR 625, 677 [285] (Bennett, Greenwood and Besanko JJ): 'intellectual' effort is protected, not just mechanical effort or investment.

141 Copyright's communicative function requires an internal balance and this should limit the scope of its rights. This internal balance can be seen embedded in copyright law. Consider – to take a more obvious example – how the idea/expression distinction embodies the need for some part of expression to be controlled, with other parts left free for reuse to enable communication. Similarly, a copy might be so transformed that it no longer 'reproduces' the original.

142 Except in the very broad sense that they capture certain relationships between pixels that constitute ideas/concepts typical of what they are described to depict, as explained by Sag, 'Copyright Safety for GenAI' (n 33) 319 and Brauneis (n 35) 31.

We argue that a teleological approach is preferable because it aligns the meaning of reproduction with its role in the copyright scheme.[143] Still, this may be a bridge too far for some Australian courts. Some case law might seem to demonstrate a preference for a technical approach to the interpretation of the exclusive rights.[144] If a court is persuaded that reproductions do occur during machine learning processes, for completeness, it is appropriate to consider here whether any transitory copies made in this process are allowable under an exception.

The most obvious path might be the temporary copying exceptions.[145] Incidental reproductions made during the machine learning process may be considered 'temporary', and thereby the exceptions in sections 43A and 43B of the *Copyright Act* could apply. In particular, section 43B exempts from infringement copies made during technical processes, for example, those stored on a computer's Random Access Memory ('RAM') and subsequently deleted after use. In contrast to the more permanent copies made for training dataset purposes, the machine learning process may involve an AI model 'copying and re-copying' the training dataset as part of its technical process.[146] In this way, section 43B may be applicable considering these are 'incidentally made as a necessary part of a technical process'. The exceptions, however, only apply where the original copy for the training dataset is held to be lawfully made.[147] Considering the likelihood that initial permanent copies are infringing, this exception is unlikely to apply for many processes of machine learning and therefore, if reproductions are found to occur, model developers would infringe copyright during this process.

As to the fair dealing provisions provided in the *Copyright Act*,[148] the exception for research and study purposes is most applicable to the activities of dataset creation.[149] It is unlikely that other fair dealing exceptions will be applicable and thus will not be considered.[150] Fair dealing for research and study purposes

143 Reading the purpose of the reproduction right differently, Matulionyte has argued for a broader reading of reproduction on the basis of functional equivalency: Matulionyte, 'Reconceptualising the Reproduction Right' (n 32). In our view, this would dissolve historical boundaries to the right that have excluded actions which experience (rather than exploit) works from copyright's ambit. We worry that such an approach takes copyright too close towards a tort of unfair competition, a path the High Court of Australia has decided not to follow.

144 For example, in determining who is responsible for making a copy, the Full Federal Court has considered that this could be both the user who pressed 'record' instigating the making of a copy, and the cloud TV service provider who retained possession, ownership and control of the physical copy made on its hard disc (even though this resulted from a wholly automated process): *National Rugby League Investments Pty Ltd v Singtel Optus Pty Ltd* (2012) 201 FCR 147, 165 [67] (Finn, Emmett and Bennett JJ). In *Australian Video Retailers Association Ltd v Warner Home Video Pty Ltd* (2001) 53 IPR 242, the exclusive right to make a 'copy' of a cinematograph film in section 86(a) was interpreted as requiring there be a specific point at which an article or thing can be identified in which visual images and sounds are embodied, which would mean the embodiment of tiny sequential fragments of a film stored in the RAM of a DVD player or PC whilst it is played would not count as a relevant 'copy': at [63].

145 *Copyright Act* (n 35) ss 43A, 43B.

146 Matulionyte, 'What Are the Options?' (n 8) 428–9.

147 *Copyright Act* (n 35) s 43B(2)(a)(i).

148 Ibid ss 40, 41, 41A, 42, 43.

149 Ibid s 40.

150 Ibid ss 41, 41A, 42, 43; Matulionyte, 'What Are the Options?' (n 8) 427.

may apply where use is for purely academic (ie, non-commercial) research, for example, by academic institutions.[151] Here, the appropriate question is whether the party responsible for creating the dataset does so for their own research or study purposes, as opposed to it ultimately being used by others for such purposes.[152] Creating an image dataset for the purposes of training an AI model in a university or academic environment, solely to undertake a research goal, may satisfy the test. However, datasets created by academic institutions for research and study purposes may then be used by external parties for commercial purposes, and this would not be covered by fair dealing provisions.[153]

In assessing the fairness of a dealing, courts consider several non-exhaustive factors. In the case of section 40 (fair dealing for the purposes of research or study) these are specified in the Act to include: 1) the purpose and character of the dealing; 2) the nature of the work; 3) the possibility of obtaining the work within a reasonable time at an ordinary commercial price; 4) the effect of the dealing upon the potential market; and 5) the amount and substantiality of the part copied.[154] Despite differences in the emphasis and applicability of the fairness factors across the fair dealing exceptions (and the non-exhaustive nature of the list), they can be seen as roughly indicative of the considerations that courts would usually bring to bear on the question of fairness.[155]

Regarding the first factor, where datasets are created by university institutions or academics for use in training models, this would weigh favourably in establishing fair dealing.[156] In consideration of the second factor, it will be more difficult to establish fair dealing as most artistic works arguably display significant creativity.[157] For the third factor, courts will have to consider the market and availability of licensing in Australia.[158] This element is unclear. From current international cases, it is evident that licensing is available to GenAI developers and could have been negotiated, yet a consensus could not be reached on the price of such agreements.[159] For factor four, courts may wish to account for the downstream

151 Matulionyte, 'What Are the Options?' (n 8); Weatherall (n 56).

152 *De Garis v Neville Jeffress Pidler Pty Ltd* (1990) 37 FCR 99 ('*De Garis*').

153 Ibid 105–6; *Haines v Copyright Agency Ltd* (1982) 42 ALR 549.

154 *Copyright Act* (n 35) ss 40(2)(a)–(e). Section 113E (fair dealing for purpose of access by persons with a disability) lists these factors but excludes the possibility of obtaining the work within a reasonable time at an ordinary commercial price.

155 The section 40(2) list was drafted to implement the recommendations of the Franki Committee and based on principles derived from fair dealing case law. As such, the Copyright Law Review Committee has suggested that it is 'reasonable to assume' that these matters are also relevant for the determination of the fairness of a dealing for purposes other than research and study: Copyright Law Review Committee, *Simplification of the Copyright Act 1968: Part 1: Exceptions to the Exclusive Rights of Copyright Owners* (Final Report, September 1998) [4.09]. This report was cited by the ALRC in their *Copyright and the Digital Economy* (Issues Paper No 42, August 2012) [246], noting at footnote 302: '[later], at [6.36], the CLRC also referred to comments to similar effect made by Professors Ricketson and Lahore in each of their loose-leaf services'.

156 *Copyright Act* (n 35) s 40(2)(a).

157 Ibid s 40(2)(b).

158 Ibid s 40(2)(c).

159 *Getty (US) v Stability AI* Complaint (n 3).

effects of GenAI models in disrupting the market of creative industries.[160] Whilst pure academic research may not detrimentally affect the market, datasets that are then sold for use in commercial AI models may disrupt the market for human creativity and thus cause detriment.[161] Finally, the fifth factor requires analysis of the amount and substantiality of the work copied.[162] As most or all of a work is likely to be used, a court would look unfavourably upon this when determining fair dealing.[163] Balancing these factors, it is unlikely that a court would find fair dealing considering the likely commercial nature of models and especially in light of the potential impact of their use on the creative industries.

In summary, unless willing to embrace a teleological reading of reproduction, courts are likely to consider 'copying' occurring during the machine learning process to infringe the reproduction right of copyright holders. The most likely exception to this would be situations where the dataset is encoded into no more than probabilistic representations of the original images in latent space vectors via machine learning where the argument that a *reproduction* is made is weaker. No exceptions are likely to permit infringing acts of copying by the dataset creators and model developers in the vast majority of cases.

C Foundational Models

After the machine learning process is complete, the result is a trained GenAI model – often referred to as a 'pre-trained' or 'foundational' model – that can then be further fine-tuned using additional data or machine learning processes.[164] In respect to model fine-tuning (which requires machine learning), the analysis above applies. An AI model trained on input data is not simply a 'magic box' or a copy of its training dataset, so further consideration is needed as to whether infringing 'copies' can be considered to be stored within the final AI model.[165] This section will examine the question of whether a resulting trained AI model *itself* is

160 Matulionyte, 'What Are the Options?' (n 8) 424.

161 Ibid. A theory of market dilution has begun to be examined in US courts. In *Bartz v Anthropic* (n 125) the plaintiff's claims of market dilution were rejected by the court (they alleged that 'training LLMs will result in an explosion of works competing with their works'): at slip op 17 (Alsup DCJ). In *Kadrey v Meta*, by contrast, Chhabria J found in favour of Meta, dismissing the plaintiff's claims of infringement, but intimated a good chance of success if with further evidence of the market dilution effects had been presented, stating '[no] matter how transformative LLM training may be, it's hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books': *Kadrey v Meta Platforms, Inc*, (ND Cal, No 23-cv-03417-VC, 25 June 2025) 23. It is unclear how an Australian court would approach this question.

162 *Copyright Act* (n 35) s 40(2)(e).

163 Matulionyte, 'What Are the Options?' (n 8) 425. This aspect could be attenuated by considering use of the entirety of the work as an essential element of machine learning – just as it is difficult to use only part of a photograph: see *Fraser-Woodward v BBC* [2005] 3 All ER 613 (although this is a UK case concerning the application of fair dealing for criticism and review and incidental inclusion exceptions). Here an argument might also be made that de-intellectualised or non-expressive uses are not a quantitatively or qualitatively substantial taking.

164 The infringement implications of fine-tuning should be considered according to the analysis in Part III(B) above.

165 Murray (n 14); *Getty (US) v Stability AI* Complaint (n 3).

an infringing copy of protected works used in its training.[166] This is significant as online repositories currently exist that enable users to access open-source models as 'foundation models', and may in effect be seen to distribute copies of works for further use and training.[167]

The parties involved at this stage of the GenAI process are usually the model developers and owners (or licensors).[168] Generally, a model developer is responsible for creating the AI model's algorithm and undertaking machine learning, and a model owner (or licensor) is responsible for the final model and its deployment.[169] Whilst it is possible that these parties may be distinct human or corporate actors, they are often the same party and therefore, the liability of the parties will be examined concurrently. Although the process of creating and ingesting training datasets may in many cases be '[reproducing] in material form' existing artistic works (as explained above in Parts III(A) and III(B)),[170] the final trained AI model itself likely does not do so. A crucial question that requires consideration is whether the final AI model embeds 'copies' of inputted works.

For model owners or developers to infringe the exclusive rights of owners of copyright in works used as training data at this stage of the process, it must be established that protected works are 'reproduce[d] in material form' within the GenAI model.[171] As outlined above, a work is reproduced where 'it is converted into or from a digital or other electronic machine-readable form'.[172] Although the training dataset is input for machine learning, it does not follow that the final trained model itself is a 'reproduction' of its training dataset as a result of the machine learning process.[173]

Thus far in the US litigation, courts seemed to have assumed AI models might contain reproductions of their training data. In Getty's case against Stability AI, a US court accepted the plausibility that an AI model could qualify as an infringing copy of its training dataset in pleadings.[174] In *Andersen v Stability AI*, Orrick J considered the storage of training data in the model as 'algorithmic or mathematical representations' not to create an impediment to the claim that 'copies' exist in the model in an infringing way.[175]

---

166 A UK court recently decided this was not supported on the evidence in *Getty Images 2025* (n 4) [592]–[602] (Smith J). See also, *Copyright Act* (n 35) ss 14, 31(1)(b)(i), (iii).

167 Particularly in the Australian context, where foundational model training will likely occur internationally, it is interesting to consider whether these models can be seen as reproducing copyright protected works domestically. There are currently over 300,000 machine learning models available for download and use on HuggingFace.co, including Stable Diffusion.

168 See Lee, Cooper and Grimmelmann (n 10) 297. Sometimes other individuals will be involved.

169 Ibid; Murray (n 14).

170 *Copyright Act* (n 35) s 31(1)(b)(i).

171 Ibid.

172 Ibid s 21(1A).

173 Ard (n 98) 517.

174 *Getty (US) v Stability AI* Complaint (n 3) [9].

175 *Andersen v Stability AI Ltd*, 744 F Supp 3d 956 (ND Cal, 2024) ('*Andersen v Stability AI 2024*'). In contrast to our analysis, and a recent UK decision (see n 166 above), Lee, Cooper and Grimmelmann (n 10) reach the conclusion that a model is an infringing copy of its training data 334–5. The reasoning for this, more thoroughly defended in Cooper and Grimmelmann (n 103), rests on a functional analysis based on the specificities of the US definition of reproduction. More importantly it relies on the assumption

Yet, as Kimberlee Weatherall explains, a final trained AI model is not like a filing cabinet – it does not contain precise copies of images from the training dataset within the model itself to then reproduce at the user's command.[176] Rather, a GenAI model can be more accurately described as a 'mathematical representation' of its training dataset: it performs statistical analysis to find patterns or clusters and stores data points of these probabilistic correlations.[177] Once the training data and algorithm have been chosen, a computer will process data for an extended period of time, with the result consisting of two files.[178] These include a 'run file' that enables the model to function when called upon and a larger file of numerical 'weights' or 'parameters' that are used by the model to distinguish key concepts.[179] As such, when a user asks a model to generate an image of a cat, to take an example used by Andres Guadamuz, the model does not refer to files of cat images stored within its training dataset and instead refers to the relationships between pixels it has learned are more likely in images with cats.[180]

Fundamentally, what is stored within an AI model are the statistical insights or parameters gathered from the training dataset, rather than 1:1 copies of the training data.[181] The mathematical information derived during model training is distinct from the image file stored in the training dataset as its purpose is to form building blocks upon which new works can be output.[182] In this way, where there are instances in which an output appears to 'copy' materials within its training data,[183] this can be explained by 'memorisation' rather than actual 'copies' within the model itself.

Cooper and Grimmelmann define memorisation as a phenomenon that occurs when '(1) it is possible to reconstruct from the model (2) a (near-)exact copy of (3)

---

that if any training data can be regurgitated then it must have been memorised, asserting that it follows a fortiori that a copyright-relevant copy must be contained within the model. As this Part explains, we do not consider that this necessarily follows, as a matter of law, in the Australian context. It is also worth noting that they consider that the most *useful* models will contain little memorisation. Thus, even if on their view an AI model is an infringing copy, it is an infringing copy of only a small number of memorised works in its training data (not all of them). It would seem extremely complex to identify which works these would be, with the most usual method of investigation being by experimentation.

176 Weatherall (n 56); Ard (n 98) 519–20.

177 Isaac Yael Sandiumenge Torres, 'Copyright Implications of the Use of Generative AI' (LLM Thesis, Universitat Pompeu Fabra, 2022–3) 14 <https://doi.org/10.2139/ssrn.4531912>.

178 Nuno Sousa e Silva, 'Are AI Models' Weights Protected Databases' *Kluwer Copyright Blog* (Blog Post, 18 January 2024) <https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/>.

179 Ibid.

180 Guadamuz (n 74) 114; Sag, 'Copyright Safety for GenAI' (n 33) 319, 323.

181 What is gained from the training dataset and stored within an AI model are embeddings or coordinates to a point in the latent space with upwards of 300 dimensions. From these coordinates in a model's latent space, the model is able to decode the appropriate features for an image output: Guadamuz (n 74) 117; Rättzén (n 94) 191.

182 Sag, 'Copyright Safety for GenAI' (n 33) 322.

183 *New York Times v Microsoft* (n 3); The New York Times Company, 'Complaint', Complaint in *The New York Times v Microsoft Corporation, OpenAI Inc* (SD NY, No 1:23-cv-11195, 27 December 2023) [2] ('*New York Times* Complaint').

a substantial portion of (4) that specific piece of training data'.[184] 'Memorisation' is a possibility with the machine learning process and has been explained as most often stemming from the inclusion of too many similar or duplicate images within a training dataset, resulting in model 'overfitting'.[185] Commentators disagree about the extent to which memorisation is an inherent part of machine learning (as well as whether it is a 'bug' or a 'feature').[186] It is generally understood to be a relatively infrequent occurrence, but good data here is lacking particularly as user output data is rarely publicly available. In the pleadings made by the New York Times as plaintiff in their case against OpenAI, it is claimed that as AI models are able to exhibit instances of 'memorisation', they thereby 'repeat large portions of materials they were trained on'.[187]

Whilst memorisation may result in highly similar outputs, this does not negate the fact that the model generates its outputs from latent concepts or information derived from works, rather than actual 'copies' of them within the model. A 'latent concept' is inferred or estimated from a series of observable variables.[188] They are inferred from their relationships with other observable or measurable variables rather than being directly observed or measured themselves.

As suggested above, in the Australian context, an argument that a 'reproduction' of works occurs in the model could be structured around the changes to the definition of 'material form' that arose following the implementation of the Australia-US Free Trade Agreement in 2004 clarifying that material form includes storage of a work, even where the work cannot subsequently be reproduced from the form of storage.[189] This would be a mistake. David Brennan argues that the results of the 2004 reforms ought not to be overstated.[190] In particular, he notes that these reforms affected neither the requirement for physical reproductions or copies, nor the necessity for a substantial part to be reproduced – both key gatekeepers for

184 Cooper and Grimmelmann (n 103) 143–4 note imprecision in how this term is employed. They distinguish memorisation from 'extraction', 'regurgitation' and 'reconstruction'. They argue memorisation inevitably occurs in all models and cite research suggesting some memorisation might be desirable for the effective functioning of the system. They explain that memorisation can, in effect, cause 'regurgitation' of the training data (defined by them as occurring when the user 'generates a (near-)exact copy, regardless of the user's intentions') and they consider this to allow the characterisation of a model as a 'copy' of its training data in copyright terms (in the US context). They take this approach by reasoning that where a model can regurgitate a substantial portion of its training data this means that the training data is 'in the model'. Though, Cooper and Grimmelmann also acknowledge the fact that memorisation of works is present in the model may not even matter to US courts – instead courts could choose to treat expression 'memorised but not regurgitated' as fair use: at 175.

185 See, eg, Guadamuz (n 74) 122; Cooper and Grimmelmann (n 103) 141; Rättzén (n 94) 220–3; Nicholas Carlini et al, 'Quantifying Memorization Across Neural Language Models', (Conference Paper, International Conference on Learning Representations, 6 March 2023) <https://doi.org/10.48550/arXiv.2202.07646>; Stella Biderman et al, 'Emergent and Predictable Memorization in Large Language Models' (31 May 2023) <https://doi.org/10.48550/arXiv.2304.11158>.

186 Cooper and Grimmelmann (n 103) 143–4.

187 *New York Times* Complaint (n 183) [80].

188 Sag, 'Copyright Safety for GenAI' (n 33) 316.

189 The object of this reform was to address situations like that which occurred in *Stevens v Kabushiki* (n 123), where the High Court found that transitory storage in a computer's RAM was not a 'material form': Brennan (n 76) 151–2.

190 Brennan (n 76) 152.

establishing copyright infringement. On the face of the current publicly available information, it appears unlikely that GenAI models store and reproduce copies that represent a substantial part of works present in their training data. To hold otherwise requires quite a stretch of language when it comes to what constitutes a 'reproduction' and a 'copy'. These key concepts deserve further analysis.

What does it mean to *reproduce* a work in copyright terms? This term is not defined in the *Copyright Act*, but there are limits to what can be borne by natural language and undoubtedly to how this term should be understood in light of the history and purpose of the scheme embodied in the Act. At a minimum, there must be some similarity or resemblance for a reproduction to occur.[191] In *Motorola Solutions v Hytera*, 'reproduction' was recently expressed to require:

(a) Objective similarity between a substantial part of the original work and the alleged reproduction; and
(b) A causal connection between the works such that the original work must have been copied.[192]

Perram J expanded on the limits of this concept with reference to an analogy comparing an architect's plans for an office tower to a complete set of instructions given to various tradespersons who actually put the office tower together. In his view, 'in no universe of discourse can a large number of such low level instructions be seen to be a reproduction of the architect's plans'.[193] Indeed, the requirement for some resemblance for there to be a copyright-relevant 'copy' was an obstacle to the recognition of digital forms of works, overcome by the addition of section 21(1A) to the *Copyright Act*.

The limits of the natural language meaning of 'reproduction' explain the need, for example, for special provisions in section 21 deeming that reproducing object code from source code (and vice versa)[194] or two-dimensional reproductions of three-dimensional artistic works (and vice versa), are relevant reproductions. These versions do not bear direct resemblance to each other. Yet, they are more closely representative of one another than data stored in an AI model is to its training data. An AI model contains statistical information *about* the contents of its training data – not reproductions of it.[195] Indeed the model is *very significantly smaller* in

191 Sam Ricketson and Christopher Creswell, *Law of Intellectual Property: Copyright, Design and Confidential Information* (Thomson Reuters, rev ed, 2024) [9.145].

192 *Motorola Solutions, Inc v Hytera Communications Corporation Ltd* [2022] FCA 1585 [1188] (Perram J) ('*Motorola Solutions v Hytera*').

193 Ibid [1189]. There may be a reproduction where following very precise instructions can only lead to an exact reproduction, as occurred in relation to regulatory guidance around the appearance of kiwi fruit packaging in *Plix Products v Frank M Winstone (Merchants) Ltd and Ors* (1986) FSR 63 ('*Plix Products*'). This situation can be distinguished. Whilst the instructions in that case were intended to precisely and exactly guide the later creation of kiwi packaging, the information or 'instructions' extracted from images in machine learning has a more generic and undefined role to be used as building blocks for the creation of new images.

194 Perram J describes this as creating a 'statutory' fiction that there is a reproduction: *Motorola Solutions v Hytera* (n 192) [1202].

195 *Purefoy Engineering Co Ltd v Sykes Boxall & Co Ltd* (1955) 72 RPC 89, 99 (Evershed MR) explaining the need to avoid an over-expansive understanding of copying by taking (cited with approval in *Desktop Marketing Systems Pty Ltd v Telstra Corporation Ltd* (2002) 119 FCR 491 [118] (Lindgren J)).

file size terms than the training data it derives from.[196] It seems unlikely that an Australian court would find that a model encodes copies of its training works in a copyright relevant sense, strengthening the argument that the model itself is not a 'copy' of the training data.[197]

Assuming this hurdle can be surpassed, whether this process can result in an infringing copy of the training data will also depend upon the courts' interpretation of whether a 'substantial part' of a protected work is taken.[198] To satisfy the 'substantial part' test, courts have found that 'a vital or material part' or an 'essential part of the work' may be enough, even if it is a small quantity of the work itself.[199] A substantial part must evidence some of the intellectual effort that made the work original, and hence protectable, in the first place. In the context of GenAI, the information gathered about the artistic works in the training dataset and stored in the model has been described as 'factual data', with commentators doubting that it would contain enough of the original expression from the analysed data to be infringing.[200]

Following this line of thought, a court may turn to the idea-expression dichotomy to assist in determining whether the statistical information drawn from a work can form a 'substantial part' of it.[201] As a general principle of international copyright law, the idea-expression dichotomy is recognised in international agreements to which Australia is party, including article 9(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights ('TRIPS') and the Guide to the Berne Convention.[202] In this way, copyright has been confined to protecting the particular expression created by the copyright owner, rather than information in it. There are reasonable grounds to find that an AI model records no more than facts about, or ideas from, its training data (mathematical or statistical information pertaining to an image's features in the sense of criteria indicated by probabilistic relationships between pixels).

In line with points made above, Australian courts may find the concept of 'non-expressive' use helpful in shedding light on whether information retained in an AI model constitutes a 'substantial part' of its training data.[203] As specified,

---

196 Cooper and Grimmelmann (n 103) 187–8, n 178.

197 Cf *New York Times* Complaint (n 183) [80]. See the next Part for a discussion of the well-known characters issue.

198 *Copyright Act* (n 35) ss 10(1), 31(1)(b)(i); *SW Hart & Co Pty Ltd v Edwards Hot Water Systems* (1985) 159 CLR 466.

199 *EMI Songs Australia Pty Ltd v Larrikin Music Publishing Pty Ltd* (2011) 191 FCR 444, 491 [188] (Jagot J) ('*Larrikin*'); *Hawkes & Son (London) Ltd v Paramount Film Service Ltd* [1934] Ch 593, 606–7 (Slesser LJ).

200 Michael W Carroll, 'Copyright and The Progress of Science: Why Text and Data Mining Is Lawful' (2019) 53(2) *University of California Davis Law Review* 893, 954.

201 *Copyright Act* (n 35) ss 10(1), 31(1)(b)(i).

202 *Marrakesh Agreement Establishing the World Trade Organization*, opened for signature 15 April 1994, 1867 UNTS 3 (entered into force 1 January 1995) annex 1C ('*Agreement on Trade-Related Aspects of Intellectual Property Rights*') art 9(2) ('*TRIPS*'); World Intellectual Property Organization, *Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971)* (WIPO Publication No 615(E), 1978); Torres (n 177) 10.

203 Sag considers that courts in the US have concluded that acts that do not communicate the 'original expression' to a new audience – namely, a 'non-expressive' use – will not impede upon the interest copyright aims to protect: Sag, 'Copyright Safety for GenAI' (n 33) 306. He makes this point

the information extracted and present in an AI model's latent space after training can be characterised as mathematical data pertaining to the works of the training dataset and not an expression of the works themselves.[204] What is retained is not elements of the work *as a* copyright work. The mathematical information or parameters stored in the AI model should be likened to factual insights from the works, as opposed to a 'substantial part' of the original expression.[205] Therefore, even if (contrary to our view) a court did find reproductions of training data exist in the AI model, there is likely no taking of a substantial part of the original works.

In summary, by one route or by another, a court would likely find that there is no reproduction in material form of a substantial part of original works within a trained model. Therefore, model owners and developers do not infringe upon the reproduction right of copyright owners and subsequently, no copyright exceptions need to be examined.[206] Similarly, as copies are not contained in models, there is no infringement of the communication to the public right when models are provided in Australia, such as occurs, for example, on Hugging Face (an online platform where developers can share and download pre-trained AI models).

D Output Stage (Prompting AI Models and Generating Images)

This section considers whether infringement can be established at the AI output stage. Whilst Part III(C) above explained that it is likely that there are no reproductions of a substantial part of protected works that exist inside trained AI models, there is continued uncertainty as to whether AI outputs 'reproduce' their training data in a relevant sense.[207] If they do, this may raise questions of liability for both end users and model developers. Outputs 'objectively similar' to protected works might also infringe the right to communicate a work to the public.[208] In our analysis, we find challenges establishing that there is a 'reproduction' of a 'substantial part' of a protected work and complications around attributing liability. The most significant and underappreciated hurdle to a finding of infringement, however, is the potential lack of 'causal connection'.

We begin to explore the question of where AI outputs infringe by following the path usually taken in the scholarship (which tends to skip over the causal connection problem by assuming a relevant derivative link exists between outputs

---

notwithstanding the application of the 'fair use' exception in that context. In a recent pre-publication version of a report released by the US Copyright Office, it appears to suggest that AI models do make expressive use of copyright works on the basis that their outputs have expressive characteristics: United States Copyright Office, *Copyright and Artificial Intelligence Part 3: Generative AI Training* (Pre-publication Report, May 2025) 47–8. See also criticism of this report: Edward Lee, 'Opinion: Why the Copyright Office's "Pre-publication" Report is Flawed: Both Procedurally and Substantively', *ChatGPT Is Eating the World* (Blog Post, 12 May 2025) <https://chatgptiseatingtheworld.com/2025/05/12/opinion-why-the-copyright-offices-pre-publication-report-is-flawed-both-procedurally-and-substantively/>.

204 Guadamuz (n 74) 117.

205 See, eg, *Baigent v Random House Group Ltd* (2006) 69 IPR 143 ('*Baigent*'), in which a UK court stressed that infringement would not occur where only factual aspects had been copied from a work of speculative historical fiction.

206 *Copyright Act* (n 35) ss 10(1), 31(1)(b)(i), 40–43B.

207 Ibid s 31(1)(b)(i).

208 Ibid s 31(1)(b)(iii).

and inputs). We begin by asking whether an exclusive right reserved to the copyright owner is infringed in AI output. Following the teleological approach suggested above, we question whether AI outputs really 'reproduce' protected expression in a relevant sense. We ask whether potentially infringing acts could be said to be done in respect of a 'substantial' part of a protected work. On this point, we discuss the problem of recognisable characters – an example of creative work easily reproduced via prompting.[209] We highlight factors that might influence the attribution of responsibility to the model developer or the user if 'objectively similar' outputs were held to be communicated to the public or reproduced in AI output (contrary to our view).

After observing that infringement is by no means clear even following the typical path, we go on to explain the fundamental flaw underlying any infringement theory established via this route: the absence of a causal connection. Analysing the legal issues via the supply chain approach allows us to expose the error of assuming a derivative link exists. This is because it demonstrates a potential break in the chain of causation at the previous stage (models do not contain relevant 'copies' of training data) which may negate a finding of infringement. We conclude by explaining how evidentiary presumptions that allow for a causal connection to be inferred complicate this analysis.

The usual starting point for the infringement analysis is to consider whether an act of reproduction has occurred.[210] Whilst some output may (apparently rarely) appear very similar to images within their training dataset, it is generally understood that text-to-image AI models are not designed to reproduce images in a copyright-relevant way.[211] Notably, the generation of every new output usually involves a 'random seed' which incorporates a degree of unpredictability into the AI model.[212] A study conducted by researchers Nicholas Carlini et al from Google and Princeton University using the Stable Diffusion text-to-image AI model sought to replicate images within its dataset, testing memorisation in the model.[213] Out of the 175 million possible images, they were only able to successfully replicate 109 to a degree considered 'near-copies'.[214] Although this standard of 'near-copies' may differ from a legal 'reproduction', it does indicate that it is highly unlikely

209 The issue of copyright infringement of 'characters' forms a part of the new complaints brought by Disney, Universal and Warner Bros against Midjourney Inc, the creators of text-to-image GenAI model Midjourney in the US. Claims filed by both media companies allege copyright infringement of key characters, including Scooby Doo and Superman, whilst Midjourney rejects these claims on the basis of 'fair use' of such copyrighted materials: Winston Cho, ' Warner Bros. Discovery Sues AI Giant Midjourney for Copyright Infringement in Major Legal Battle', *Hollywood Reporter* (online, 4 September 2025) <https://www.hollywoodreporter.com/business/business-news/warner-bros-discovery-sues-ai-company-copyright-infringement-1236361610/>.

210 *Copyright Act* (n 35) s 31(1)(b)(i).

211 Weatherall (n 56).

212 Lemley (n 13) 202.

213 Sag, 'Copyright Safety for GenAI' (n 35) 131, citing a preprint Nicholas Carlini et al, 'Extracting Training Data from Large Language Models' (Research Paper No 2012.07805v2, 2021) <https://arxiv.org/abs/2012.07805>.

214 Carlini et al (n 213) 5–6.

for output images to be a close match for specific images in the training dataset.[215] Replication of works in the training data would seem to occur when there are many duplicates in the training data rather than exclusively due to an end user's careful prompting. In this way, it may still seem possible to attribute a degree of liability to the model owner through the offence of authorisation where they are responsible for selecting the initial training dataset.[216]

As established in Part III(C), there may be no digital reproductions or 'copies' made or stored within the AI model itself as it only contains 'mathematical parameters' derived from statistical analysis of the training dataset.[217] In this way, a model's ability to generate new output is not predicated on reproducing copies of the training data. In contrast to the pleadings in the New York Times case referred to in Part III(C), this was acknowledged by the plaintiffs of another US case against Stability AI brought by three artists whose works were allegedly used as training data.[218] Whilst the plaintiffs admit there are no copies of protected works sitting in an AI model for use in generating output, they argue that Stable Diffusion nonetheless created 'latent images' of all 5.8 billion works that it was trained on (this would have equated to 240 terabytes in storage if the full images were stored).[219] In relation to these claims, practical understandings of the technology seem to negate the argument that there is a 'reproduction' of the images that may satisfy section 21(1A) in such 'latent images'.[220] Researchers suggest that exact 'reproductions in material form' are unlikely since approximately 240 terabytes of images cannot be stored within an AI model which is usually the size of only eight gigabytes.[221] As explained in Part III(C), it is statistical insights encoded in model weights, rather than reproductions of actual images, that enable AI models to distinguish key 'latent concepts' in the training data and generate output accordingly.[222]

Aside from the question of reproduction, the generation of any 'objectively similar' outputs may seem to infringe upon a copyright holder's exclusive right to communicate their work to the public.[223] As most text-to-image AI models are hosted online for the public to use, it may be argued that the act of generating output is an infringing act as it 'communicates' the protected work to the public.[224] In determining the party responsible, section 22(6) turns attention to the end user by providing that a communication is 'made by the person responsible for determining the content of the communication'.[225] Arguably, the end user does not merely 'click on a link to gain access' to a webpage; rather the prompting from an

215 Cf *New York Times* Complaint (n 183) [80].
216 *Copyright Act* (n 35) s 36(1A).
217 Weatherall (n 56); Sag, 'Copyright Safety for GenAI' (n 33).
218 *New York Times* Complaint (n 183) [1]–[9]; *Andersen v Stability AI 2023* (n 27).
219 *Andersen v Stability AI 2023* (n 27); Torres (n 177) 40.
220 *Copyright Act* (n 35) s 21(1A).
221 Torres (n 177) 45. See also Guadamuz (n 74) 115.
222 Guadamuz (n 74) 115. See explanation of 'latent concept' in Part III(C).
223 *Copyright Act* (n 35) s 31(b)(iii).
224 Ibid s 10(1).
225 Ibid s 22(6).

end user requires a degree of intentional effort to arrive at their desired output.[226] As such, it may seem likely that an output that is 'objectively similar' to a protected work infringes the communication right of copyright holders in a way that may be attributable to the end user who prompted it.

Following the usual line of thinking, discussions regarding copyright infringement can tend to turn to whether the output reproduces a 'substantial part' of a work in the training data.[227] Courts will assess the quality of what has been copied, reflecting on the degree to which a defendant's work makes substantial use of the features that make a copyrightable work original.[228] Still, courts will need to consider the idea/expression dichotomy and recognise that copyright will not protect the ideas or styles within an artistic work.[229] Where an end user prompts the AI model to generate an artwork by Picasso, for instance, it is likely that the generated output will be similar in style and ideas rather than a replica of an existing artwork.[230] The incorporation of unprotectable 'styles' or 'ideas' will not render an output infringing. It must be acknowledged, however, that the line between idea and expression is notoriously slippery. There are cases where taking the concept and feel of a work has been considered sufficient.[231] Recently, for example, copyright infringement was found where a company produced lookalike packaging for children's snacks, having used the plaintiff's packaging as a 'benchmark'.[232]

A special case, as Sag writes, is where the prevalence of 'memorisation' within an AI model leads to outputs that closely reflect visual characters (copyrightable under US law and an instance commonly cited as indicative of infringement in the AI context).[233] Using the example of 'Snoopy', Sag argues that AI models 'learn' the relationship between the word 'Snoopy' and the features that make this character original and recognisable.[234] Even where an output does not exactly replicate an existing Snoopy artwork, it may be possible for courts to deduce that a recognisable rendition of Snoopy is 'objectively similar' to the original character.[235] US courts might find that output images are an improper appropriation of existing works within a dataset, especially in the case of widely recognisable characters.[236] But this example is less of a concern in Australia where a broad US-style notion

226 Ibid s 22(6A).
227 Ibid s 14.
228 Ibid; *Larrikin* (n 199).
229 *Baigent* (n 205); *Designers Guild Ltd v Russell Williams (Textiles) Ltd* [2000] 1 WLR 2416; *Madden v Seafolly Pty Ltd* (2014) 313 ALR 1.
230 Lee, Cooper and Grimmelmann (n 10) 340.
231 *Elwood Clothing Pty Ltd v Cotton On Clothing Pty Ltd* (2008) 172 FCR 580 ('*Elwood*'). Analysing the complexity of this issue in the US context: Benjamin LW Sobel, 'Elements of Style: Copyright, Similarity and Generative AI' (2024) 38(1) *Harvard Journal of Law and Technology* 49.
232 *Hampden Holdings IP Pty Ltd v Aldi Foods Pty Ltd* [2024] FCA 1452.
233 Sag, 'Copyright Safety for GenAI' (n 33) 330.
234 Ibid.
235 Sag, 'Copyright Safety for Gen-AI' (n 33) 327–35.
236 This is alleged in new cases initiated against Midjourney and Chinese AI company, MiniMax, by media firms, including Disney, Warner Bros, and Universal: Cho (n 209).

of character copyright is not recognised.[237] It would seem to be uncommon for an AI output to resemble a specific individual image of a character such as Snoopy (which would be the protected artistic work here).

In cases of 'objective similarity' between a protected work and generated output, the liability of both the end user and the model owner should be distinguished.[238] As discussed above, where an end user directly inputs a prompt that results in output that may appear to be 'objectively similar' to a training data image, a court may consider the end user to be primarily liable.[239] As copyright infringement is effectively a strict liability offence, the requisite intention and specific wording of the prompt created by the end user likely will not be critical.[240]

Assessing the liability for a model owner requires a different investigation into their *authorisation* of infringing acts. A court may place weight on the statutory factors in section 36(1A) of the *Copyright Act*, including any control that could be exercised by the model developer and steps taken by them to prevent the infringing output.[241] By analogy, an AI owner may be likened to an internet service provider in that it is a user who ultimately prompts the AI model to infringe copyright, such that it does not have significant power to prevent the infringing act.[242] The presence of some 'memorisation' in an AI model may not strengthen a finding of authorisation as model owners arguably do not have the power to prevent it entirely. Still, they may need to show that they have taken steps to reduce memorisation, for example, by removing duplicates in training datasets and/or ensuring that any implementing GenAI system is designed to reduce this phenomenon.[243] Where model developers have not allowed artists to voluntarily 'opt-out' of using their works for model training, it is possible that this could be taken into account to support a finding of authorisation.[244]

But, as indicated at the beginning of this section, perhaps much of the foregoing is moot. Underlying this analysis, a crucial question arises that is insufficiently considered in the existing Australian scholarship.[245] The technical aspects of the

237 Jani McCutcheon, *Literary Characters in Intellectual Property Law* (Edward Elgar Publishing, 2023) 31–2 <https://doi.org/10.4337/9781788114325>.

238 *Roadshow Films Pty Ltd v iiNet Ltd* (2012) 248 CLR 42 ('*Roadshow Films*').

239 Ibid.

240 Australian Law Reform Commission, *Traditional Rights and Freedoms: Encroachments by Commonwealth Laws* (Report No 129, December 2015) 304.

241 *Copyright Act* (n 35) ss 36(1A)(a)–(c).

242 *Roadshow Films* (n 238).

243 *Copyright Act* (n 35) ss 36(1A)(a), (c); Sag, 'Copyright Safety for GenAI' (n 33) 338–43. Cooper and Grimmelmann (n 103) 215 note that GenAI models are embedded in systems, for example, public facing software services. These systems can be designed so that memorised outputs are controlled in several ways: on entry by filtering or modifying user prompts; in the way a model is designed to respond to prompts (its alignment); and by filtering outputs delivered to users.

244 *Copyright Act* (n 35) s 36(1A)(a).

245 Considering this point in the US context, see Lemley (n 13) 202–8; and in the UK context, see Guadamuz (n 74) 123–4 (considering that evidence of similarity should be less probative of copying in the case of GenAI); and in the EU context, see Jan Bernd Nordemann, 'EU Law: Generative AI, Copyright Infringements and Liability' *Kluwer Copyright Blog* (Blog Post, 23 January 2024) <https://copyrightblog.kluweriplaw.com/2024/01/23/eu-law-generative-ai-copyright-infringements-and-liability-my-guess-for-a-hot-topic-in-2024/>.

GenAI process, as discussed above and understood in light of the supply chain model, raise the question of whether there is a break in the chain of causation required for infringement to be established.[246] The requirement for a causal connection is a fundamental aspect at the heart of the cause of action – for there can be no *copying* in the basic sense that copyright requires without an unbroken chain of causation. Independent creation is a complete defence to any claim of copyright infringement.[247] In the classic case of *Ladbroke (Football) Ltd v William Hill (Football) Ltd*, for instance, Lord Reid stated that a reproduction 'does not include cases where an author or compiler produces a substantially similar result by independent work without copying'.[248] Without a causal connection, any action for infringement or authorisation of that infringement would fall away.

As outputs from machine learning are generated from mathematical information in the model, as opposed to any 'copies' of original works in the training data, there are grounds to argue that this is a case of independent creation. The generation of an end user's output is shaped only by the mathematical information within the AI model – at this stage, any copies within the training data have been previously deleted.[249] Consequently, a court may deduce there is a break in the chain of causation to find no instance of copyright infringement.[250] Practically speaking, this would mean that an output from an image-generating AI model is essentially non-infringing, even if it appears to be an 'objectively similar' image.[251] How likely is a court to come to this view in practice? Technical and legal understandings

246 *Creation Records Ltd v News Group Newspapers* (1997) 39 IPR 1. Lee, Cooper and Grimmelmann (n 10) 335 themselves do not consider this to be the case – they reason that if any memorised reproduction can be output, this fact indicates that it must be from a copy in the model.

247 *Francis Day & Hunter Ltd v Bron* [1963] Ch 587, 623–5 (Diplock LJ) ('*Francis Day & Hunter*').

248 *Ladbroke (Football) Ltd v William Hill (Football) Ltd* [1964] 1 WLR 273, 276 (Lord Reid).

249 On the importance of proving a causal connection: see, eg, *Autospin (Oil Seals) Ltd v Beehive Spinning* [1995] RPC 683, where an infringement claim failed because the design drawings upon which the oil seals were allegedly based were not produced in evidence. The claimant had sought, instead, to rely on drawings it was established came into existence after the manufacture of the seals. These could not support their claim.

250 Lemley (n 13) 202–3. This does not necessarily entail that an output which appears to be 'memorised' is only coincidentally similar. Indeed, the chances of this occurring are said to approach the impossible: Cooper and Grimmelmann (n 103). Independent creation is not limited to the coincidental – it also extends to situations where works appear similar because they aim to solve the same problem (eg, the design for a letterbox draught excluder in *Kleeneze Ltd v DRG Ltd* [1984] FSR 399).

251 One approach might be to construct an argument following *Plix Products* (n 193). In that case, a designer following very precise regulatory specifications for kiwi fruit packaging was found to infringe copyright in the packaging design which regulators had closely followed as a model for those specifications. This was despite the designer never having seen the initial product design because the instructions matched it so closely and, being based on the original packaging, a derivative link could be established. Machine learning can be distinguished from this example in that the latent concepts embodied in AI models represent building blocks of expression, and not precise instructions that naturally lead to the recreation of the images on which they are based. Where memorisation has occurred, this hurdle may be more easily surpassed. Whilst the degree of memorisation may differ between GenAI models, as Cooper and Grimmelmann (n 103) explain, it is generally agreed that the most useful models *generalise* from their training data, rather than *memorising* it.

of GenAI are constantly improving, but the data used to develop popular GenAI models is still shrouded in obscurity and protected as trade secrets.[252]

In this context, where evidence is unclear, a court's first port of call might be to rely on evidentiary presumptions of copying. This complicates matters. In *Francis Day & Hunter Ltd v Bron* ('*Francis Day & Hunter*'), it was held that causal connection could be inferred with evidence of objective similarity and access, establishing a rebuttable presumption of copying that eases the practical burden on plaintiffs seeking to establish an infringement claim.[253] Where 'objective similarity' exists between a GenAI output and a protected image, 'access' can be established by the existence of that image in the training dataset, making up one of the original inputs to the AI model. Of course, the often private nature of an AI model's training dataset would be a hurdle here. If a claim that a relevant dataset has been created through indiscriminate web scraping (as many contend has often been the case) is accepted by a court, there is a possibility that the mere existence of a work on the internet could be enough to establish access for these purposes. In this way, the presumption of copying may provide a means to 'fudge' the causal link necessary to establish infringement.

It would then fall to a defendant to establish independent creation. On the one hand, in the case of model developers, this would provide a valuable incentive to disclose technical details about the machine learning process. On the other hand, it would be challenging for a user who has prompted the model to obtain evidence that might rebut this presumption. Reliance on evidentiary presumptions might appear pragmatically desirable in allowing for a remedy where there may appear to be a harm (ie, in the case of a highly imitative AI output that serves as a market substitute for an original work). Yet, this approach is likely to result in a significant degree of uncertainty for market participants. Furthermore, where a causal link can be inferred, it will still fall to the courts to distinguish whether a 'substantial part' of a work has been taken where, on the facts of the case, an output is seen as potentially similar to an existing work. As discussed above, this may be a challenging hurdle to surpass.

IV IMPLICATIONS OF THE DOCTRINAL ANALYSIS

There are several ways in which the preceding doctrinal analysis may feel unsatisfactory. The first is that should litigation occur, its result is unlikely to reveal the *general* copyright position. Any outcome would necessarily be dependent on the technical characteristics of the specific model used; the particularities of the AI

252 Sag, 'Copyright Safety for GenAI' (n 33); Foong (n 32); Matulionyte, 'Exception, Compensation or Both' (n 32); Martin Senftleben, 'Generative AI and Author Remuneration' (2023) 54(10) *International Review of Intellectual Property Law and Competition Law* 1535 <https://doi.org/10.1007/s40319-023-01399-4>; Lee, Cooper and Grimmelmann (n 10).

253 *Francis Day & Hunter* (n 247) 616 (Lord Diplock). This inference of copying is recognised in Australia, for example, in *Pacific Gaming Pty Ltd v Aristocrat Leisure Industries Pty Ltd* (2001) 116 FCR 448, [19] (Sackville, Finn and Kenny JJ).

outputs complained of; and the ability of the party being sued to provide evidence that can rebut a presumption of copying; amongst other factors. Plaintiffs will need to think carefully about the stage in the supply chain that they are targeting.[254] In the face of legal uncertainty that cannot be easily resolved in court, technology companies are likely to continue to move fast without regard to whether they may break things.

A second issue is that whether machine learning infringes copyright may depend on the specific technical means used. It is worth reflecting upon whether, as a matter of principle, the process of machine learning ought to be considered infringing. Consistency might be achieved by appealing to the argument that machine learning should not be seen to reproduce protected expression in a copyright-relevant manner via a teleological reading of 'reproduction'.[255] This approach has advantages insofar as it links the default concept of reproduction to copying that involves a communicative use or taking the expressive value of a work beyond the ideas, facts or raw information contained within it.[256] As some have stressed, in the analogue era, copyright did not prevent people from reading or appreciating a work, or drawing on ideas from it to learn.[257] To adopt a default understanding of reproduction that encompasses similar or even lesser technical acts undertaken with 'copy-reliant' technology shifts an important balance at copyright's heart between what is protected and what is free.[258]

Uncertainty as to whether a court would adopt such a teleological approach to interpreting 'reproduction' may provide an argument for legislative clarification by way of an exception to permit the 'non-expressive' or technical uses of works in GenAI, the path recently recommended by the Productivity Commission.[259] Here,

---

254 Note that the case evolved on these grounds in the German LAION case: *Kneschke* (n 5). For a nuanced explanation of the complexities of the AI supply chain and the impact of choices across it for a copyright infringement analysis in the US context: Lee, Cooper and Grimmelmann (n 10).

255 Sag, 'Copyright Safety for GenAI' (n 33); Sag, 'Copyright and Copy-Reliant Technology' (n 35).

256 Sag and Yu (n 92).

257 Ibid 7, noting

> [i]n the 1990s, for the first time, we began to see economically significant acts of copying that had nothing to do with communication or transmission of the underlying expression … [These uses] pose a difficult conceptual question for copyright law: the centerpiece of copyright law is the exclusive right to reproduce the work, yet the purpose underpinning that right is to allow the author to control the communication of his or her original expression to the public while still allowing ideas, facts, abstractions, and artistic methods to be freely copied.

258 Sag, 'Copyright and Copy-Reliant Technology' (n 35). It is worth noting that in the recent pre-publication report at United States Copyright Office (n 203) 48, the US Copyright Office distinguishes machine learning from human learning on the basis that

> [h]umans retain only imperfect impressions of the works they have experienced, filtered through their own unique personalities, histories, memories, and worldviews. Generative AI training involves the creation of perfect copies with the ability to analyze works nearly instantaneously. The result is a model that can create at superhuman speed and scale.

This does not alter the position that machine learning involves retaining or copying unprotectable aspects of copyright works. For a critical discussion of non-expressive use questioning scholarly accounts appealing to this concept in the GenAI context: see Brauneis (n 35) 22–45.

259 *Productivity Commission Report* (n 7), although the Government is not currently inclined to adopt it: see n 24 above. Margoni and Kretschmer (n 137) 689 argue that 'in a properly designed copyright framework there should be no need for a TDM exception, as the extraction of factual information from protected

lessons can be drawn from jurisdictions that have already adopted exceptions for computational data analysis of works and 'text and data mining' ('TDM'),[260] although it must be remembered that these were typically drafted without the specific challenges of GenAI in mind.[261] The UK has a narrowly drafted TDM exception confined to non-commercial uses.[262] The EU combines a broad exception for certain non-commercial research uses (and users) with a more complex exception for commercial uses that is subject to rightsholder opt-out provisions.[263] Japan has recently adopted one of the broadest exceptions, covering uses for a 'non-enjoyment purpose'.[264] This acknowledges works as acts of communication to an audience, but it is not unlimited because it also accounts for 'unreasonable prejudice' to the copyright owners' interests. Singapore, too, has recently enacted an exception for computational data analysis drafted broadly, alongside a 'fair use' provision similar to that of the US.[265] Embedding these two provisions is rare and shows Singapore's decision to provide a favourable jurisdiction for AI development.

A well-drafted exception for non-expressive use could help sustain a GenAI industry in Australia and provide the conditions necessary to minimise problems of bias in the development of local AI models. It is beyond the scope of this article to evaluate the benefits and disadvantages of such an exception and the range of approaches to its drafting, but this is an area that would benefit from

---

content is external to the remit of copyright'. Sag, 'Copyright Safety for GenAI' (n 33) 306, citing Lemley and Casey (n 33) 782. Sag, a prominent proponent of the non-expressive use line of thinking prefers to accommodate it at the point of exceptions (in the US context where fair use provides a flexible balancing tool).

260 On the differences: see Sag and Yu (n 92) 14–15.

261 Nicola Lucchi and Serra Hunter, 'Generative AI and Copyright: Training, Creation, Regulation' (Study, Policy Department for Justice, Civil Liberties and Institutional Affairs, July 2025) 33 consider the view the EU TDM exception permits GenAI training to have arisen 'through a combination of textual ambiguity, regulatory silence, and widespread industrial reliance'. Note also that many only apply if access to protected works was 'lawful', which could limit their use where copyright works are accessed by spoofing accounts, CAPTCHA bypassing or where anti-bot counter-measures are evaded: Kathy Bowrey et al, Submission No 141 to Productivity Commission, *Harnessing Data and Digital Technology Interim Report* (9 September 2025) <https://doi.org/10.26190/unsworks/31625> ('*Submission to Productivity Commission*').

262 *Copyright, Designs and Patents Act 1988* (UK) s 29A. There have been calls to expand this exception to all uses, however these are still under consideration: *Productivity Commission Report* (n 7) box 1.7.

263 *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC* [2019] OJ L 130/92 arts 3, 4. See Margoni and Kretschmer (n 137).

264 *Copyright Act 1970* (Japan) art 30(4). As such, Japan's TDM exception allows use of works to the extent that the purpose does not involve enjoying ideas or emotions in the work, evidently recognising the value exchange in the communication of expression of works: Craig, 'AI-Copyright Challenge' (n 129) 152. The Japanese Copyright Subdivision of the Cultural Council has also clarified two areas pertaining to the exception. First, this exception is qualified by consideration that the use must not unreasonably prejudice the interests of copyright holders, meeting the Berne Convention's three-step test. The Council also outlined that the use of copyright works for AI training is not covered by the exception where licences are available in the marketplace: Legal Subcommittee under the Copyright Subdivision of the Cultural Council, Japan Copyright Office, *General Understanding on AI and Copyright in Japan* (Overview, May 2024) 5–10.

265 Sag and Yu (n 92) 28.

further research.[266] As AI-friendly jurisdictions emerge, calls for international harmonisation on AI regulation will likely grow louder. Yet, with competition between nations seeking to become AI superpowers growing, it seems unlikely that an international treaty will emerge soon. Instead, Matthew Sag and Peter K Yu have predicted global convergence and a 'race to the middle', with nations responding to the challenges of AI training by adopting exceptions that will broadly converge but that can also account for local considerations.[267]

A third unsatisfactory conclusion is that our analysis tends to suggest that copyright can more effectively target the input stage than the output stage. This is problematic as companies will likely seek to shield themselves from liability by creating datasets and undertaking machine learning in favourable jurisdictions (which will then be imported).[268] More importantly, it does not touch the point of sale, where significant revenue may be generated for AI companies. This has led some to argue for pragmatically expanding the concept of 'reproduction' to rule out the interpretation that we have recommended.[269] There are many reasons why a bright line on infringement might be sought. For example, copyright infringement could underpin a licensing-based solution to compensate creators whose works have been used in the creation and refinement of these useful and profitable tools.[270] Or, it could underpin a levy-based approach targeting AI systems that are capable of serving as substitutes for human artistic creativity.[271] It seems to offer hope as an anchor for schemes to ensure the potentially significant revenue expected to accrue to companies developing GenAI tools is shared with the creators whose works were essential in building them.

Before launching into such schemes, however, it is worth considering the damage that might be done if they are built on a distortion of fundamental copyright principles (even if this is to achieve the laudable aim of supporting creators). Copyright rests on more than a vague and general idea of misappropriation.[272]

---

266 For more: see ibid. A strong case for even more ambitious reform to copyright in response to the challenges of the digital environment was made by the ALRC in its *Copyright and the Digital Economy Report* (n 36) in 2013. Any reform should consider the copyright system as a whole and its interaction with related bodies of law: see Bowrey et al, *Submission to Productivity Commission* (n 261).

267 Sag and Yu (n 92) 45–6.

268 Matulionyte 'What Are the Options?' (n 8). Thereby, the avenues for fair financial compensation for copyright holders will diminish regardless of Australia's stance on AI model inputs: Weatherall (n 56).

269 Matulionyte, 'Reconceptualising the Reproduction Right' (n 32) argues that reproduction should extend to machine learning on the basis that the information retained in an AI model is the 'functional equivalent' of a reproduction in economic and technical terms. Pragmatic expansion risks creating an unwieldy ad hoc body of jurisprudence as has arguably occurred with the right to communicate to the public in the EU.

270 See, eg, Gervais et al (n 43); Alain Strowel, 'ChatGPT and Generative AI Tools: Theft of Intellectual Labour?' (2023) 54(4) *International Review of Intellectual Property and Competition Law* 491, 494 <https://doi.org/10.1007/s40319-023-01321-y>.

271 This approach is preferred by Senftleben (n 252), drawing on a concept of the collective rights of the community of authors or *domaine public payant* developed by Adolf Dietz. Considering a statutory remuneration right: see Christophe Geiger and Vincenzo Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' [2024] 52 (April) *Computer Law and Security Review* 105925:1–19 <https://doi.org/10.1016/j.clsr.2023.105925>.

272 *Victoria Park Racing* (n 139). Alfred C Yen, 'Brief Thoughts About If Value/Then Right' (2019) 99(6) *Boston University Law Review* 2479, 2480. Dilan Thampapillai, 'If Value, Then Right? Copyright and Works of Non-human Authorship' (2019) 30(2) *Australian Intellectual Property Journal* 96.

The High Court of Australia has recently cautioned against an 'if value/then right' approach that rests on an amorphous idea of free-riding.[273] Collapsing copyright into a generalised protection against unfair competition risks depriving it of conceptual tools that have been refined for many years and developed incrementally via the common law method. If there is any hope of certainty increasing over time and successful adaptation to technological change, it is important that courts continue to apply notions such as 'reproduction', 'causal link' and 'substantial part' in a principled manner with dual goals of doctrinal coherence and consonance with copyright's normative aims in mind.[274]

We have argued that courts may need to reach for a normative lens to help better understand what a copyright-relevant 'reproduction' ought to encompass. It is worth mentioning that copyright law is underpinned by a mix of different theoretical justifications.[275] Government reports and discussions around policymaking often adopt an instrumental economic lens.[276] In the most simplified form, these see copyright as an incentive for the creation and dissemination of creative work. Despite the prevalence of economic approaches, moral and natural rights arguments that conceive of copyright as a reward or recognition for the contribution of labour or personality are also prominent.[277] As technology provides new challenges, scholars are looking beyond the 'stale' instrumentalist vs naturalist theory binary.[278] Increasingly, scholars and courts are embracing broader perspectives,[279] arguing for a pluralist justificatory toolbox,[280] or defining mid-level principles shared across theoretical disagreements.[281] More research is needed to update theoretical approaches to copyright so that they can provide

273 *IceTV* (n 140).

274 If policymakers are persuaded that machine learning should be regulated by copyright, this ought to be clarified by extending the definition of 'reproduction' to encompass another deemed reproduction, as was done in relation to source code and object code, or three-dimensional reproductions of two-dimensional art works, etc.

275 William Fisher 'Theories of Intellectual Property' in Stephen R Munzer (ed), *New Essays in the Legal and Political Theory of Property* (Cambridge University Press, 2001) 168.

276 Productivity Commission, Australian Government, *Intellectual Property Arrangements* (Inquiry Report No 78, 23 September 2016) 55–8.

277 Rebecca Giblin, 'A New Copyright Bargain? Reclaiming Lost Culture and Getting Authors Paid' (2018) 41(3) *Columbia Journal of Law and the Arts* 369, 372: '[t]here is no single coherent normative rationale for copyright'.

278 Jeremy Sheff, 'Philosophical Approaches to Intellectual Property Scholarship' in Irene Calboli and Maria Lillà Montagnani (eds), *Handbook of Intellectual Property Research: Lenses, Methods and Perspectives* (Oxford University Press, 2021) 295: 'normative theory of IP has long been trapped in a stale debate over a false dichotomy between (often vulgar) utilitarianism and (often cramped) Lockeanism'.

279 Bowrey et al (n 46) 106–16. Carys J Craig, 'Critical Copyright Law and the Politics of "IP"' in Emilios Christodoulidis, Ruth Dukes and Marco Goldoni (eds), *Research Handbook on Critical Legal Theory* (Edward Elgar Publishing, 2019) 301; Deidre A Keller and Anjali Vats, 'Bridging Race + IP: The Challenges and Potential of Utilizing Transdisciplinary Methods to Undo the Unbearable Whiteness of Intellectual Property' in Irene Calboli and Maria Lillà Montagnani (eds), *Handbook of Intellectual Property Research: Lenses, Methods and Perspectives* (Oxford University Press, 2021).

280 See, eg, Justin Koo, 'A Justificatory Pluralist Toolbox: Constructing a Modern Approach to Justifying Copyright Law' (2020) 42(8) *European Intellectual Property Review* 469.

281 See, eg, Robert Merges, *Justifying Intellectual Property* (Harvard University Press, 2011).

clearer guidance in the light of new challenges posed by GenAI in the Australian context. We suggest, as a minimum, that any such view should understand the role of copyright in fostering a particular system of (human) public communication.

In a creative ecosystem up-ended by GenAI, courts will be called upon to more clearly articulate and uphold the normative values that copyright aspires to serve.[282] This is likely to occur before the slow wheels of academia turn to address the complex theoretical issues. It may appear tempting to pragmatically pick a winner that copyright should 'incentivise' – whether that be creators of works used in training data, the creative industries disrupted by GenAI, or the new tech pioneers – and then stretch copyright concepts to serve their ends. In our view, this is not the best way forward for copyright's development.

The problem cuts deeper than scepticism about whether courts are best placed to make this call. Fundamentally, a purely instrumental economic incentive view of copyright misses key dimensions of authorship as a social and creative practice.[283] A single-minded emphasis on the economic interests of copyright owners (or users for that matter), provides a blinkered view of the balance needed between the rights of artists to create and communicate their works, the public interest in accessing them, and the progression of human culture.[284] What is at stake transcends discussions about the calibration of incentives that drive market actors. In interpreting aspects of the test for copyright infringement that are stretched to their limits by AI, we have argued for a focus on teleological interpretations guided by copyright's purpose to sustain the creation and communication of human creative expression by protecting creators, but also by ensuring sufficient access to the ideas and raw material necessary for creativity in its varied forms.

This leads us to perhaps the most profound way in which the doctrinal analysis in the previous section may feel unsatisfactory. That is the conclusion that despite what has been previously assumed, the correct outcome in many cases may be that GenAI models and AI output do not infringe copyright in images used in training data. GenAI promises to upend the creative industries and has been seen to represent a serious challenge to the livelihoods of creative people. Yet, these tools are only as valuable as they are because of their use of pre-existing creative work. The breadth, richness and creativity of that work is the bedrock of the value proposition upon which GenAI tools rest. In this context, the call to arms and

282 Craig, 'AI-Copyright Challenge' (n 129) 135. Rebecca Giblin and Kimberlee Weatherall, *What If We Could Reimagine Copyright?* (Australian National University Press, 2017) 10 <https://doi.org/10.22459/WIWCRC.01.2017>.

283 Craig, 'AI-Copyright Challenge' (n 129) 29. Jessica Silbey, *The Eureka Myth: Creators, Innovators, and Everyday Intellectual Property* (Stanford University Press, 2014) <https://doi.org/10.1515/9780804793537>. It also cannot account for an artist's innate motivation to create: Radha Pull ter Gunne, 'Copyright Theory and Musical Creation: Oh, Why Is There Such a Disconnect?' [2020] (27) *University of New South Wales Law Journal Student Series* 1, 15–16.

284 Craig, *Copyright, Communication and Culture* (n 125) argues for a relational theory of copyright law that views copyright as focused on promoting the *communication* of expression between people.

the appeal to copyright law is understandable.[285] How can copyright *not* protect vulnerable creators? Surely, they are its *raison d'être*.

In a recent article, Craig warns of the dangers of falling into the 'AI-copyright trap', that is, the view that copyright law is the best regulatory tool to support human creators and culture in the face of AI challenges. Relying on copyright to protect creators from AI's impact, she argues, is misguided because the arguments to do so rely upon fallacies such as the view that value inherently warrants property rights or that all unauthorised copying is wrong.[286] The schemes proposing to use copyright to redress the 'AI problem', she reasons, are unlikely to result in compensation reaching smaller artists and actual creators as the benefits will be hoovered up by larger players and intermediaries.[287] Further, deploying copyright's concepts in a pragmatic, rather than principled, manner risks backfiring with further copyright expansion that does not serve the interests of human creativity, playing into the hands of corporate actors.[288] Copyright law, she contends, may not be the right place to look for answers to the disruption that will be caused by this technology.[289]

She is not alone in coming to this conclusion.[290] Micaela Mantegna argues that the harms of GenAI should be understood in the broader context of AI ethics, with better and more holistic solutions to be found outside the context of copyright law.[291] This is not to say that copyright reform should not occur to promote certainty and desirable policy outcomes.[292] But rather, that when approaching this question, policymakers should consider the limits of what copyright can accomplish.[293]

---

285 Creator concerns are not confined to loss of direct revenue or even the substitution effect of AI outputs. They also extend to reputational harms. Some of these concerns could be captured by passing off or consumer law (such as the circulation of works 'in the style of' a particular creator that might be associated to them without being authored by them). Moral rights might provide some relief although their application is uncertain: Rita Matulionyte, 'Can AI Infringe Moral Rights of Authors and Should We Do Anything about It? An Australian Perspective' (2023) 15(1) *Law, Innovation and Technology* 124 <https://doi.org/10.1080/17579961.2023.2184138>.

286 Noting that the copyright regime does not restrict all types of copying – it leaves space for the copying of ideas, style, works in the public domain: Craig, 'The AI-Copyright Trap' (n 2).

287 Ibid. For instance, the proposed settlement in the US case of *Bartz v Anthropic* (n 125) would provide $1.5 billion to members of the class action – which equates to roughly $3,000 USD per work used. Despite the historic terms of settlement, the award for each individual artist still raises the question of what is 'adequate' in these scenarios, particularly where impacts are seen to future livelihoods and reputation.

288 Ibid 108: '[i]t is a trap in the sense that it may satisfy the wants of a small group of powerful stakeholders, but it will harm the interests of the more vulnerable actors who are, perhaps, most drawn to it. Once entered, it will also prove practically impossible to escape.'

289 She points instead to a range of other legal and policy tools that can help address the various harms AI may cause: ibid.

290 Micaela Mantegna, 'ARTificial: Why Copyright Is Not the Right Policy Tool to Deal with Generative AI' (2024) 133 *Yale Law Journal Forum* 1126.

291 Mantegna suggests alternative compensation and redistribution methods like universal basic income, taxation, government incentives, and culture funds: ibid 1131.

292 It is not within the scope of this article to detail or defend any specific reform measure. Any such proposal should rely upon a fuller discussion of the normative basis underpinning such a proposal, reflect engagement with the views of a broad range of stakeholders and be developed in light of the available evidence base. This undertaking is beyond the scope of this article, which aims to accomplish a preliminary step in such a process by providing a more accurate view of the existing legal position.

293 See also Dennis Crouch, 'Using Intellectual Property to Regulate Artificial Intelligence' (2024) 89(3) *Missouri Law Review* 781, 825–7 <https://doi.org/10.2139/ssrn.5014647>. Any copyright law reform

V CONCLUSION

By focusing on new technical understandings of GenAI viewed through the lens of the supply chain model, this article has revealed that the answers to copyright infringement questions associated with GenAI – at least in the context of text-to-image GenAI models – are more complex than some previous scholarship has assumed. It is by no means clear that the GenAI process results in copyright infringement at every stage and in every case. Creating datasets using unauthorised copies of images will certainly infringe copyright. Whether their subsequent use in machine learning also infringes may depend upon the specific technical means used and a court's willingness to embrace a teleological understanding of 'reproduction'. The technical complexities of this process – even within the confined context of text-to-image GenAI models – and the lack of transparency from technology companies about their data collection processes presents serious challenges in clearly determining the legal position on this point.

Currently available information seems to suggest that AI models themselves may not contain infringing copies of works captured in training datasets. Further, analysing the process in this supply chain method reveals that this may be a link that breaks the chain of causation necessary to establish infringement in relation to model outputs. There is considerable uncertainty as to whether this could be overcome by the presumption from *Francis Day & Hunter* and, in the case of users of GenAI, whether they could access evidence needed to rebut this presumption. In addition, the latter stages of the AI supply chain, representing machine learning, model refinement, and deployment, present challenges in relation to whether a 'substantial part' of copyright protected source material is 'reproduced' in a relevant sense. The answer will be uncertain because it is dependent on the facts and judicial interpretation of these concepts in a new context.

In the final section, we discussed some of the implications of this doctrinal analysis, identifying several ways in which it may feel unsatisfactory. Legislative reform of some sort may be on the horizon but at this stage it is difficult to tell what shape it will take. Copyright law can be an enabler or a brake on the development of GenAI. In determining the role of copyright in the era of GenAI, policymakers should be guided by our shared interest in encouraging creativity, supporting creators, and ensuring human cultural development in the context of technological and economic advancement.[294] GenAI has a potential role in encouraging a more participatory culture by providing accessible tools for creative expression.[295] Beyond the text-to-image generation example discussed here, GenAI tools promise impressive gains in innovation and efficiency that might be harnessed in the public

---

should be undertaken in a joined-up manner, considering the complex interactions between copyright and other bodies of law (e.g. privacy, consumer protection and competition law): Bowrey et al, *Submission to Productivity Commission* (n 261).

294 See Giblin and Weatherall (n 282) 18 discussing how the 'public interest' might be used as a benchmark for reimagining a better copyright system.

295 Craig, 'The AI-Copyright Trap' (n 2) 24. Where infringement liability extends to the *users* of these new tools, a chilling effect on the creation and communication of creative expression may occur.

interest (as well as for private gain). To make good on these promises, these tools will need to be built using good quality training data.

At the same time, concerns that the proliferation of GenAI outputs that mimic the artistic works of creators has the potential to displace the incentives for human creation and effectively substitute humans within the marketplace of expression are certainly real.[296] This can be seen in the escalating policy tug of war over copyright protections between creative industries and tech companies. As a result, it is important for any reform agenda to respond meaningfully to the varied disruption that GenAI will bring to the human creative ecosphere – both economic and non-economic. It must not be forgotten that this disruption will have costs as well as benefits. Market concentration, competition, and transparency concerns are also significant issues. The balance to be struck is not straightforward.

The best way for copyright law to remain relevant is to stay true to its normative heart by placing special value on the many and varied forms of human creativity and communication.[297] Any copyright reform will need to fit within a framework of measures needed to regulate the broader risks and harms AI may entail.[298] Although copyright is now in the spotlight thanks to the flurry of international litigation, it is not the only tool in our regulatory toolbox.[299] Creators are not the only people whose livelihoods will be threatened by pervasive adoption of AI. As AI threatens to displace human workers in many fields of endeavour, there is a need for ambitious policy action capable of addressing the full range of labour issues that will arise. This must occur in the context of a broader conversation about the way in which we value, remunerate, and distribute different kinds of work – including creative work.

296 Senftleben (n 252) 1535. On the need for a broader framing in relation to intellectual property rights more generally: Peter S Menell, 'Mapping the Intellectual Property/Social Justice Frontier' in Steven D Jamar and Lateef Mtima, *The Cambridge Handbook of Intellectual Property Law and Social Justice* (Cambridge University Press, 2024) 21–2.

297 Overseas scholars are increasingly making the point that there is a need for a richer notion of the public interest in deeper social and cultural communication. Senftleben (n 252) 1549; Martin Senftleben, 'AI Act and Author Remuneration: A Model for Other Regions?' (Research Paper, 19 March 2024) <https://doi.org/10.2139/ssrn.4740268>; Margaret Chon, 'Relational Innovation and the Public Benefits of Copying' (2024) 39(3) *Berkeley Technology Law Journal* 1169 <https://doi.org/10.2139/ssrn.5069862>.

298 Mantegna (n 290).

299 In many ways, current copyright law is an imperfect tool. Creators will not always be owners of copyright in their work. In many contexts the lion's share of the benefit from the copyright in creative work does not find its way to creators themselves, who may lack the bargaining power to secure significant returns in negotiations with distributors and other intermediaries. Indeed, one concern around licensing as a solution to the AI problem is that the profits from such schemes may not find its way, in relevant quantities, to those individual creators most disrupted by this technology.