Learning to Copy [GenLaw 2024]

Anonymous

June 2024

1. Introduction

This paper aims to clarify the discussion of copyright infringements associated with generative AI, which I refer to as Media Models, or MMs. It focuses on English and Canadian law.

I identify six distinct acts of copying associated with MMs:

1. Pre-training copies. These are the copies of works used in training a model.

2. Training copies. These are the copies of works used to trigger model weights when training a model.

3. Model copies. These are copies of works that are embedded in the weights of a trained model.

4. Direct prompt copies. These are copies of works made in order to prompt a model. For example, a user might prompt an LLM with a poem and the instruction to add another verse to the poem.

5. Indirect prompt copies. These are copies of works that are created by a system in order to prompt a model. For example, a user might submit a photograph and a request for it to be reimagined in the style of a new artistic work. In response to this prompt, the system might then search the internet, find works in that style, copy those works, and use those copies to prompt a MM to respond to the user's request.

6. Output copies. These are copies of works that a MM generates in response to a prompt.

I then analyse who makes each of these copies, and whether an exemption from copyright infringement liability would be available in each circumstance.

2. The copies

A copyright is *prima facie* infringed when a person causes a substantial part of a copyrighted work to be reproduced. This definition has three components.

First, a substantial part of the copyrighted work must be present in the new work. Whether the parts co-occur is a factual question. What counts as "substantial" is a normative question that mixes fact and law. Outside the digital realm, these questions of cooccurrence and substantiality are normally evaluated experientially: courts compare how a person would experience the copyrighted work versus the new work. When the new work is digital, however, it cannot be directly experienced since it is encoded as a series of bits. It must be decoded to be experienced. Nonetheless, an encoded copy of a work (such as a digital photograph of a picture) is no less a copy in its encoded form than when the copy is decoded.

Second, the new work must be made by copying the copyrighted work. This element requires the new work would not contain the substantial part that is the same between them but for the existence of the copyrighted work. Independently-created new works are not copies. It also requires that the copy be created by direct interaction between a person and a copy (or the original work itself), recursively defined.

The final element of copyright infringement is that the infringer must voluntarily (but not necessarily intentionally) have caused the creation of the new work. I discuss this matter more in the next section.

That a copy exists may sometimes be controversial. I argue that the trained weights of a model can themselves include copies of the works the model is trained on, essentially because training a model is a kind of data compression. Output copies are more ambiguous: they could be copies or they could be independent creations, but I suggest the burden of showing independence should rest on the ostensible copier. Finally, building again on *Football Dataco*, I argue that if D2 trains its model on synthetic data from D1's model, then can result in infringing the copyrights that D1 infringed, as well as infringing D1's own database rights.

3. Who makes the copies

Who makes a copy is sometimes obvious. Pre-training, training, and model copies are plainly made by the developer of a MM (or a subcontractor). Direct prompt copies are plainly made by the user.

Other times, less so. Who makes an indirect prompt copy

or output copy is more complex, and depends on how one attributes actions.

The original case to consider how responsibility should be allocated for copies caused via code is *Religious Tech Center v Netcom On-Line Comm*, [*Netcom*], from the Northern District of California. In *Netcom*, the court refused to find the operator of an online bulletin board and its ISP liable for copyright infringement, despite users of the bulletin board uploading copyrighted material and the ISP transmitting that material. The court held that liability depended on "some element of volition or causation which is lacking where a defendant's system is merely used to create a copy by a third party". The core of *Netcom* is that the volitional conduct of the defendant — deploying code "necessary to having a working system for transmitting Usenet postings to and from the Internet" — is insufficiently proximate to the copying. The court instead rested legal responsibility for the copying solely on the user who uploaded the copyrighted material.

The continuing relevance of *Netcom* is debatable, not only because it is an American case, but also because both the British and the Canadian Parliaments have added safe harbours to their copyright legislation to ensure that hosting third party content or acting as an ISP (the conduct in *Netcom*) is infringing only in limited circumstances. The texts of these safe harbours do not mention causation or volition, but, as I argue in the paper, they can be understood to clarify the framework.

Taking these conditions as more general guidance, volitional conduct that reasonably foreseeably leads to the creation of a copy does not constitute copying when, at a minimum,

1. That conduct is necessary to enable efficient non-infringing processes that are of value to society; and
2. Another person more specifically intends creating the copy.

The transmission and caching safe harbours suggest a further requirement of neutrality toward the content. Volitional conduct involving the exercise of (non-technical) judgment when creating a copy **is** sufficiently proximate.

A caveat to this framework is that there is one UK case (*Football Dataco Ltd v Sportradar GmbH*, wherein a user was held (albeit in *obiter*) to make a copy even though they did not know the actions they did (clicking on a link on a website) would have that consequence. I argue this case was wrongly decided on that point.

Setting that caveat aside, I suggest that output and indirect model copies will only be made by the user if the user appeared to intend that a copy be made. If not, then output copies should be held to be made by the creator of the model, because the creator will have made non content-neutral decisions in deciding what content to train their model on, how to finetune their model, and how prompts will be responded to.

4. Exemptions

The exemptions worth considering are fair dealing, data analysis, temporary copying, incidental use, and a licence. None exempt the aforementioned copiers from the potential liability associated with commercially available MMs.

In Anglo-Canadian law a fair dealing exemption applies only if the copying fits in both a fair dealing category and the copying is fair. In Canada, the 'research and private study' category is the most promising. Canadian courts have found that a 'research purpose' can include personal, non-scientific consumer research, such as deciding which songs are worth buying, as well as commercial research. And, more, Canadian courts have held that the category applies based on how the copy is used, not why it is made. In the UK, none of the categories are promising because even research and private study must be non-commercial. This exemption would cover only certain output copies.

The data analysis and temporary copying exemptions are both of limited relevance. The former exists only in the UK, and applies only where the "sole" purpose is non-commercial, so it wouldn't apply to commercial MMs. The latter would seem to apply to training copies, but only if this training does not facilitate a later infringement.

The incidental use exemption has some promise for model copies in Canada and output copies in both jurisdictions. It has no relevance for model copies in the UK because a model is probably a kind of compilation or database, which is therefore classed as a literary work and excluded from this exemption. It has potential relevance for model copies in Canada, but only potential. At a per-use level, every single copyrighted work may well be incidental to the function of the MM, but collectively, they are essential. For output copies the question is simpler, and can be assessed based on the purpose of the output.

The final exemption is a licence. Many works used to train MMs are found by CommonCrawl, a service that obeys the Robots Exclusion Standard (RES). The argument for licensing goes that by failing to block CommonCrawl with RES, a website owner consents to the copying. This argument must fail. First, A website owner cannot grant a licence if they do not already have the power to sublicence. Second, even where they have that power, silence should not construed as consent, unlike with search indexing. Search indexing returns a benefit to the website owner and is notorious: training an LM does not.

References

Copyright, Designs and Patents Act 1988 (UK)

Alberta (Education) v Canadian Copyright Licensing Agency (Access Copyright), 2012 SCC 37.

American Broadcasting v Aereo, Inc, [2014] 134 S Ct 2498.

CBS Songs v Amstrad Consumer Electronics, [1988] AC 1013 (HL).

Ashdown v Telegraph Group Ltd, [2001] EWCA Civ 1142.

Associated Press v Meltwater US Holdings, Inc, [2013] 931 F Supp 2d 537.

Banks v CBS Songs Ltd, (1995) Lexis Citation 1587 (ChD).

Barrett v Universal-Island Records Ltd, [2006] EWHC 1009 (ChD).

British Leyland Motor Corpn Ltd v Armstrong Patents Co Ltd, [1986] AC 577 (HL).

Canadian Broadcasting Corp v SODRAC 2003 Inc, 2015 SCC 57.

CCH Canadian Ltd v Law Society of Upper Canada, 2004 SCC 13.

CCH Canadian Ltd v Law Society of Upper Canada, 2002 FCA 187.

Cinar Corporation v Robinson, 2013 SCC 73.

Compo Co Ltd v Blue Crest Music et al (1979), 1 SCR 357 (SCC).

Delrina Corp v Triolet Systems Inc, 2002 CanLII 11389 (ONCA).

DSC Communications Corp v DGI Technologies, Inc, [1995] 898 F Supp 1183.

Entertainment Software Association v Society of Composers, 2012 SCC 34.

The Football Association Premier League Ltd v Panini UK Ltd, [2002] EWCA Civ 95.

Falcon v Famous Players Film Co, [1926] 2 KB 474 (CA).

Field v Google Inc, [2006] 412 F Supp 2d 1106.

Football Association Premier League Ltd v QC Leisure, [2008] EWHC 1411 (ChD).

Football Dataco Ltd v Sportradar GmbH, [2013] EWCA Civ 27.

Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire, [2011] EWHC 2892 (ChD).

Francis Day & Hunter v Bron, [1963] Ch 587 (CA).

Fraser-Woodward Ltd v BBC, [2005] EWHC 472 (ChD).

Hachette Book Group, Inc v Internet Archive, (2023) 664 F.Supp.3d 370 (SDNY) .

Hutton v Canadian Broadcasting Corporation, 1992 ABCA 39.

IPC Magazines Ltd v MGN Ltd, [1998] FSR 431 (ChD).

Karno v Pathé Frères, (1909) 100 LT 260.

Navitaire Inc v EasyJet Airline Co Ltd (No 3), [2004] EWHC 1725 (ChD).

Religious Tech Center v Netcom On-Line Comm, [1995] 907 F Supp 1361.

Newspaper Licensing Agency Ltd v Meltwater Holding BV, [2011] EWCA Civ 890.

Ocular Sciences Ltd v Aspect Vision Care Ltd (No2), (1996) 1997 RPC 289 (ChD).

Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd, [2013] UKSC 18.

Rogers Communications Inc v Society of Composers, 2012 SCC 35.

SOCAN, NRCC, CMRRA/SODRAC Inc - Tariff for Satellite Radio Services, 2005-2010, [2009] CanLII 101408.

Sheeran v Chokri, [2022] EWHC 827 (ChD).

Society of Composers, Authors and Music Publishers of Canada v Bell Canada, 2012 SCC 36.

Society of Composers, Authors and Music Publishers of Canada v Canadian Assn of Internet Providers, 2004 SCC 45.

Society of Composers, Authors and Music Publishers of Canada v Entertainment Software Association, 2022 SCC 30.

Sony Music Entertainment (UK) v Easyinternetcafe Ltd, [2003] EWHC 62 (ChD).

Tamawood Limited v Habitare Developments Pty Ltd (Administrators Appointed) (Receivers and Managers Appointed) (No 3), [2013] FCA 410.

The Koursk, [1924] P 140 (CA).

Tumber v Independent Television News Ltd (ITN) & Anor, [2017] EWHC 3093 (IPEC).

Uber Technologies Inc v Heller, 2020 SCC 16.

Wade v British Sky Broadcasting Ltd, [2016] EWCA Civ 1214.

York University v Canadian Copyright Licensing Agency (Access Copyright), 2021 SCC 32.

Hutchison, Cameron, *Digital copyright law* (Toronto, ON: Irwin Law, 2016).

MacGillivray, EJ, *The Copyright Act, 1911, annotated* (London: Stevens and Sons, 1912).

Mysoor, Poona, *Implied Licenses in Copyright Law* (Oxford: OUP, 2021).

Stokes, Simon, *Digital Copyright: Law and Practice* (Hart Publishing, 2014).

Copinger and Skone James on Copyright, Gwilym Harbottle, Nicholas Caddick & Uma Suthersanen, ed (London: Sweet & Maxwell, 2021).

Ang, Steven, 'The Idea-Expression Dichotomy and Merger Doctrine in the Copyright Laws of the U.S. and the U.K.' (1994) 2:2 Intl JL & IT 111.

Arnold, R & P S Davies, 'Accessory liability for intellectual property infringement: the case of authorisation' (2017) 133:Jul Law Q Rev 442.

Balganesh, Shyamkrishna, 'The Normativity of Copying in Copyright Law' (2012) 62:2 Duke LJ 203.

Bonadio, Enrico & Luke McDonagh, 'Artificial intelligence as producer and consumer of copyright works: evaluating the consequences of algorithmic creativity' (2020) 2 IPQ 112.

Carlini, Nicholas et al, 'Extracting Training Data from Diffusion Models' (2023) arXiv.org:230113188.

Craig, Carys, 'Locke, Labour, and Limiting the Author's Right: A Warning Against a Lockean Approach to Copyright Law' (2002) 28:1 Queen's LJ 1.

———, 'Putting the Community in Communication: Dissolving the Conflict Between Freedom of Expression and Copyright' (2006) 56:1 UTLJ 75.

———, 'Technological Neutrality: (Pre)Serving the Purposes of Copyright Law' in Michael Geist, ed, *The Copyright Pentalogy: How the Supreme Court of Canada Shook the Foundations of Canadian Copyright Law* (2013), 271.

Craig, Carys J & Ian R Kerr, 'The Death of the AI Author' (2021) 52:1 Ottawa L Rev 31.

Denicola, Robert C, 'Volition and Copyright Infringement' (2016) 37:4 Cardozo L Rev 1259.

Drassinower, Abraham, 'Authorship as Public Address: On the Specificity of Copyright Vis-A-Vis Patent and Trade-Mark What Ifs and Other Alternative Intellectual Property and Cyberlaw Stories' (2008) 2008:1 Mich St L Rev 199.

Hagen, Gregory R, 'Technological Neutrality in Canadian Copyright Law' in Michael Geist, ed, *The Copyright Pentalogy: How the Supreme Court of Canada Shook the Foundations of Canadian Copyright Law* (2013), 307.

Handa, Sunny, 'Reverse Engineering Computer Programs Under Canadian Copyright Law' (1995) 40 McGill LJ 621.

Hohfeld, Wesley Newcomb, 'Fundamental Legal Conceptions as Applied in Judicial Reasoning' (1917) 26:8 Yale LJ 710.

Hudson, Emily & Paul Wragg, 'Proposals for Copyright Law and Education during the COVID-19 Pandemic' (2020) 71:4 N Ir Leg Q 571.

Katz, Ariel, 'Fair Use 2.0: The Rebirth of Fair Dealing in Canada' in Michael Geist, ed, *The Copyright Pentalogy: How the Supreme Court of Canada Shook the Foundations of Canadian Copyright Law* (2013), 93.

Lee, Jyh-An, 'Computer-generated Works under the CDPA 1988' in Jyh-An Lee, Reto M Hilty & Kung-Chung Liu, ed, *Artificial Intelligence and Intellectual Property* (Oxford: Oxford University Press, 2021).

Lemley, Mark & Brian Casey, 'Fair Learning' (2021) 99 Tex L Rev 743.

Levendowski, Amanda, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93:2 Wash L Rev 579.

Liebesman, Yvette Joy & Julie Young, 'Litigating Against the Artificially Intelligent Infringer' (2020) 14:2 Fla Int U L Rev 259.

Olah, Chris et al, 'Zoom In: An Introduction to Circuits' (2020) 5:3 Distill e00024.001.

Rae, Jack W et al, 'Scaling Language Models: Methods, Analysis & Insights from Training Gopher' (2022) arXiv:211211446 [cs].

Selbst, Andrew D & Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87:3 Fordham L Rev 1085.

Somepalli, Gowthami et al, 'Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models' (2022) arXiv:221203860 [cs].

Tushnet, Rebecca, 'Architecture and morality: transformative works, transforming fans' in Kate Darling & Aaron Perzanowski, ed, *Creativity without Law* (New York University Press, 2020), 171.

Vaver, David, 'The Copyright Amendments of 1997' (1997) 12 IPJ 53.

Knopf, Howard, *Copyright and Fair Dealing: Guidelines for Documentary Filmmakers* (Documentary Organization of Canada, 2010).