Input for Public Consultation on Copyright and Artificial Intelligence (2024)

Thank you for the opportunity to give my views and opinion on the HK Government's proposed TDM exemption when it comes to the grouping of technologies – but more importantly, the legal issues surrounding copyright and labour/economic investment practices – that fall under the category of Generative Artificial Intelligence.

I. STAKEHOLDER POSITION
II. NECESSITY OF TRANSPARENCY REQUIREMENTS
III. COPYRIGHT LIABILITY
IV. DEEPFAKE LIABILITY
V. TDM COMMERCIALIZATION CAVEAT
VI. COPYRIGHTABILITY OF GENERATIVE AI OUTPUTS
VII. TRUE VIABILITY OF OPT-OUT (vs OPT-IN)
VIII. PROPOSED SOLUTIONS
IX. CONCLUSIONS

I. STAKEHOLDER POSITION

To clarify my stakeholder position in this for the sake of transparency, it is one as a professional creative, copyright, and intellectual property holder, as well as very pragmatic early adopter of promising technologies and avid avoider of hype. As a production artist and game developer, it is always necessary to maintain a competitive edge by being at the forefront of new methods to increase efficiency in production. This goes for most of my peers in production. It is very telling that that same majority has had an extremely adverse reaction to the unveiling, adoption, and forced rollout of generative AI: That reaction is not due to any kind of desire for special protectionism to being 'under threat' of our jobs being displaced due to the nature of any fair competition or progress that happens all the time, constantly. But due to the sheer exploitative nature of the current gAI paradigm due to the very lax enforcement of copyright laws that amounts to unfair competition as opposed to fair: In short, it is more about exploitative labour practices, copyright infringement, and unfair competition, than the strawman of protectionism and fear of 'progress' that the AI ecosystem likes to peddle.

II. TRANSPARENCY REQUIREMENTS

I know that very well: My own works are confirmed to have been used in the LAION datasets to train the foundational models that StabilityAI, Midjourney, MetaAI, Flux, and every other finetuned model downstream from them based, on the datasets and derived models. This can be confirmed given StabilityAI's and LAION's commitment to being open source, and as such, they partnered with SpawningAI[1] to provide a searchable database where rightsholders can inquire about whether their data has been used for training. I was never given a chance, or choice, to opt-out. My works are being used in a way that directly competes against me economically, which goes against the entire point and purpose of having copyright protections in the first place.

This is how my peers have been able to satisfy the legal requirements to launch the Artist class action lawsuit against the defendant companies of StabilityAI, Midjourney, DeviantArt, and RunwayML. I am acquainted only superficially with one of the key plaintiffs in this case, Karla Ortiz, who is spearheading the litigation and testified in front of the US Senate last year[2] – however, despite the geographical distance, we all know and talk with each other. I in small part, had a role in helping shape the amended complaint of the legal filings – specifically the copyright infringement and inducement section: And because of our combined efforts, **all** motions to dismiss on copyright infringement & inducement grounds for all plaintiffs in the class action were denied, the litigation is now moving forward onto discovery.[3] This is why OpenAI was not part of the legal suit as one of the defendants: They have steadfastly refused to, and fought, all efforts both legislative[4] and judicially, to force them to be transparent and disclose their training data. Unlike the other gAI companies that

[1] https://spawning.ai
[2] https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_ortiz.pdf
[3] https://www.theverge.com/2024/8/13/24219520/stability-midjourney-artist-lawsuit-copyright-trademark-claims-approved
[4] https://time.com/6288245/openai-eu-lobbying-ai-act/

need to rely on external institutions like LAION for their common crawl datasets, OpenAI has done everything in house. It is only recently that anyone, anywhere, will get even a chance to query what datasets were used in OpenAI's generative AI products, and that is only because they are being forced to due to discovery requirements due to other pending litigation[5] that is being spearheaded by the same legal firm: Joseph Saveri Law Firm, LLP, that is representing my artist peers in their own class action.

These efforts, as well as the New York Time's own litigation to sue OpenAI[6], are making progress because it is easier to compare the text-based outputs directly with any plagiarized copyrighted texts for substantial similarity[7], and then launch litigation on that basis of suspicion to even be able to lead to further discovery. Whereas for anything more complicated beyond simple black and white plaintext, (Video, images, audio) where the (copyrighted) data is going to be interpolated extensively making identifying the causality between inputs and outputs to satisfy the legal substantial and/or striking similarity requirements to even use as a basis of suspicion to even begin to launch litigation or enforce any other methods of accountability for violating copyrights, extensively difficult.

The HK Government's own proposed TDM exemption and opt-out paradigm, itself would be rendered toothless without basic transparency obligation with regards to the data sets that have been used for training generative AI systems. This is a basic form of accountability that falls on generative AI companies and is no different from any other transparency obligation that various other industries face to be compliant with laws and industry standards. There is no valid excuse in the world that stands up to basic scrutiny that AI companies can deploy: If it is onerous for them (It isn't) – the party with direct access to their own data – to disclose their records, much less *engage* in record keeping, then it is ludicrous to somehow think that it is less onerous for any other party with no direct access to the gAI company's records and systems to somehow be expected to do so in lieu.

The HK Government's current position that current copyright laws are sufficient in terms of proving violations of copyright ignore the jurisprudential discrepancy introduced with its own proposed TDM exemption given the opacity of being able to even have a basis of suspicion to launch litigation to begin with: As shown in the above examples of Open Source disclosures being the reason that litigation that my peers are involved in was able to be launched to begin with, excluding OpenAI due to them being anything but 'open' with regards to their own training data and the example of how certain bad faith actors absolutely cannot be trusted to police themselves or be held to account: To balance the very nature of this TDM exemption, transparency requirements are a must to ensure that the genuine legal and economic interests of rightsholders are respected, and cannot be optional at the largess of the gAI companies. This has been recognized by the EU with their AI act which was passed late 2023 where they implemented transparency requirements[8] (That OpenAI themselves extensively fought against) as well as California – where the vast majority of incumbent gAI firms are located - where Governor Gavin Newsom within the last month signed Bill AB 2013 into law[9] requiring gAI companies to disclose their training data. Without transparency with regards to the training data that is used by gAI companies, there can be no form of accountability, and the government and rightsholders would rely exclusively on the largess and self-reporting from gAI companies that they are in compliance with copyright laws even with the opt-out paradigm. To which, even they have difficulty even commenting on if they will even comply[10].

That makes as much sense as relying exclusively on financial institutions to be honest with regards to their internals and that no external audits or disclosures are required to make sure that they are not involved in money laundering and are in compliance with anti-money laundering laws and measures. Much less on the individual or company taxpayer if they are able to independently shield their finances from a proper audit by the IRD and for the department to rely exclusively on the taxable entity to be fully honest and transparent in their declarations with no independent verification mechanism.

III. COPYRIGHT LIABILITY

[5] https://www.hollywoodreporter.com/business/business-news/openai-training-data-inspected-authors-copyright-case-1236011291/

[6] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

[7] https://nytco-assets.nytimes.com/2023/12/Lawsuit-Document-dkt-1-68-Ex-J.pdf

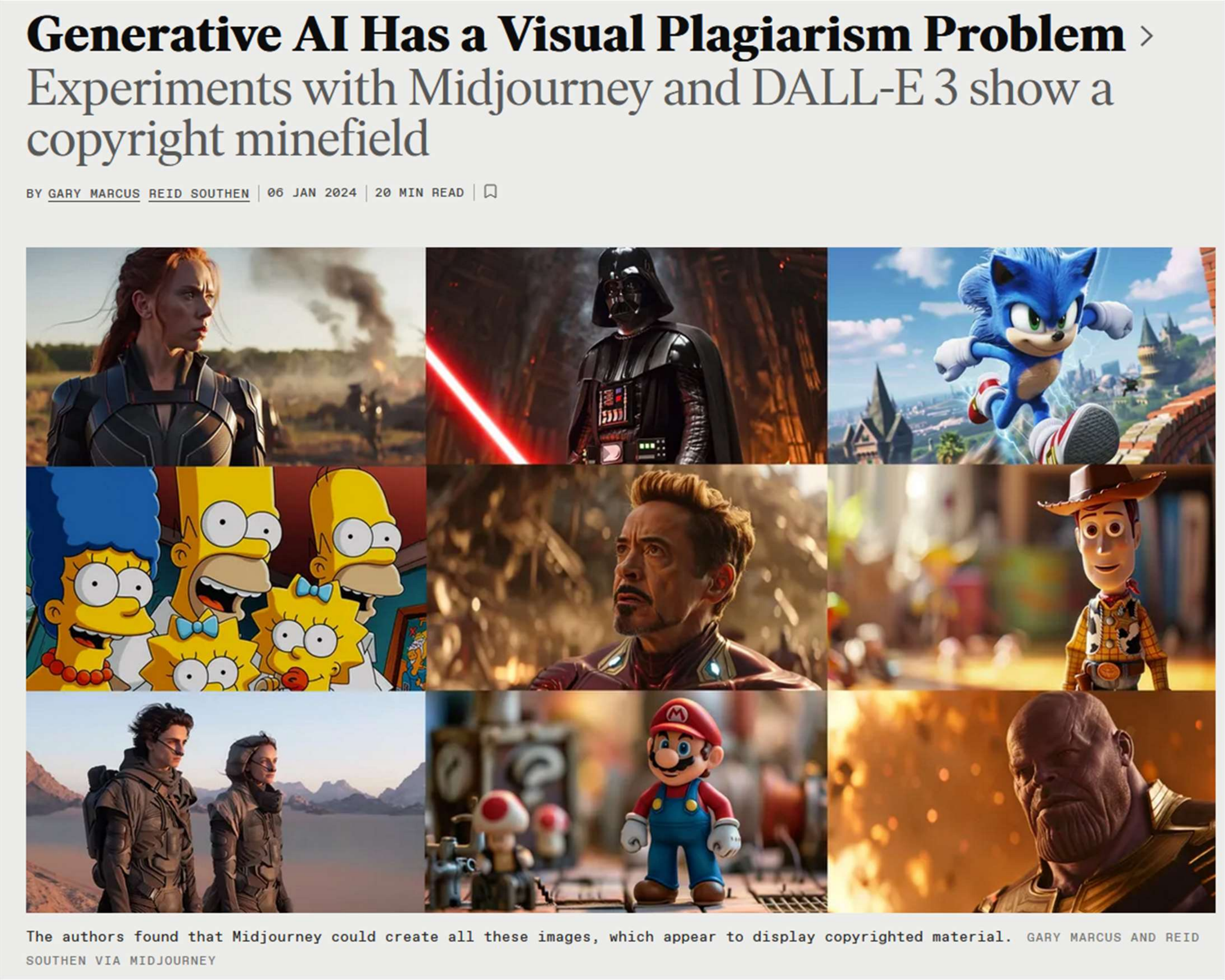
[8] https://artificialintelligenceact.eu/recital/107/

[9] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013

[10] https://techcrunch.com/2024/10/04/many-companies-wont-say-if-theyll-comply-with-californias-ai-training-transparency-law/

Liability for copyright infringement has to fall, at bare minimum, on the gAI developer as they are the one primarily at fault for doing the infringing necessarily with regards to unauthorized downloading, copying, and use of the work for gAI training, even regardless of if the prompter for the gAI output is directly soliciting copyright material or not. As it is entirely possible and highly probably from a statistical point of view that the user can generate infringing outputs without even necessarily asking for it[11] As has been extensively shown in this article Co-Authored by Gary Marcus, and my good friend and colleague, Reid Southen:

https://spectrum.ieee.org/midjourney-copyright



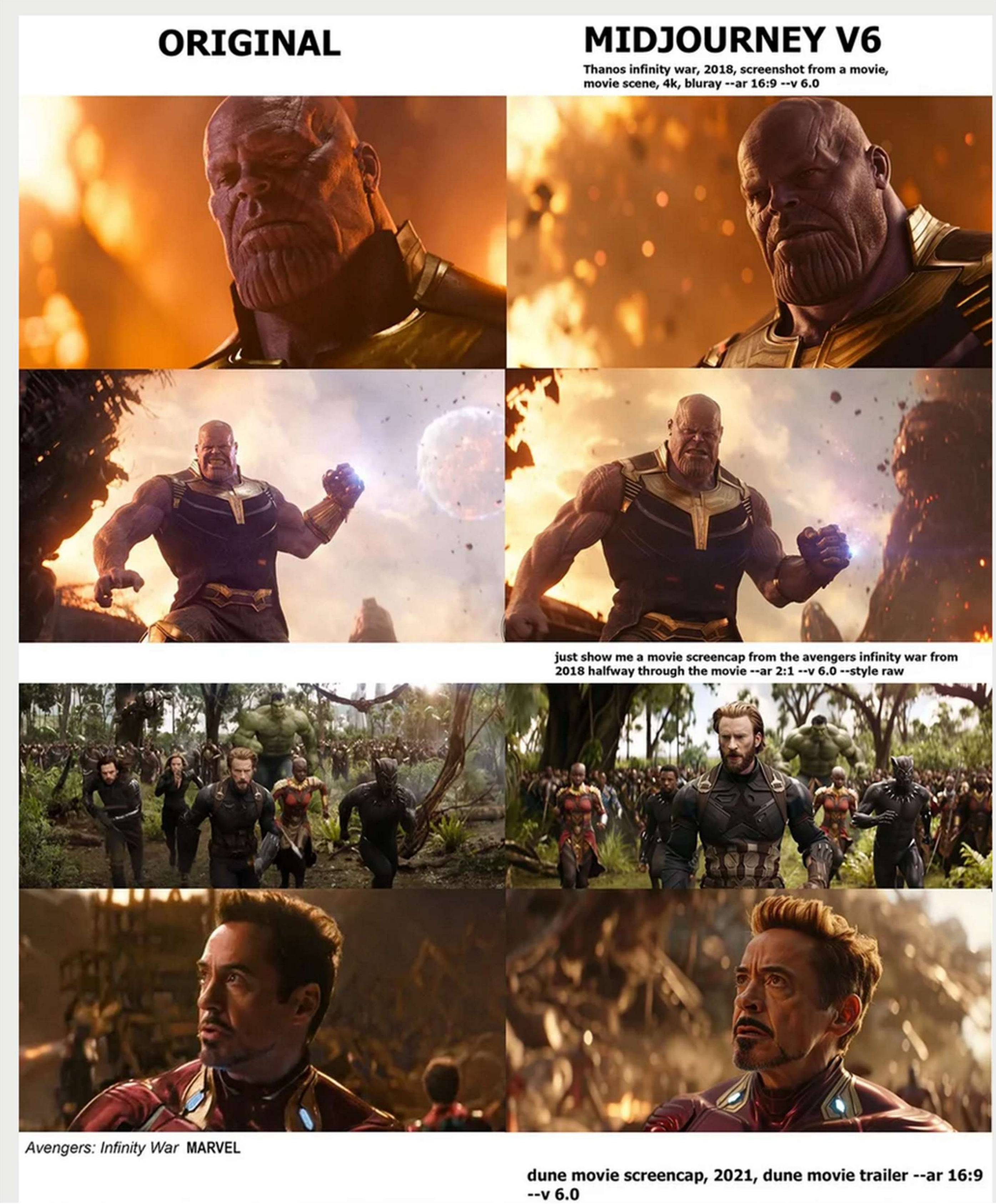
Generative AI Has a Visual Plagiarism Problem ›

Experiments with Midjourney and DALL-E 3 show a copyright minefield

BY GARY MARCUS REID SOUTHEN | 06 JAN 2024 | 20 MIN READ |

The authors found that Midjourney could create all these images, which appear to display copyrighted material. GARY MARCUS AND REID SOUTHEN VIA MIDJOURNEY

[11] https://spectrum.ieee.org/midjourney-copyright

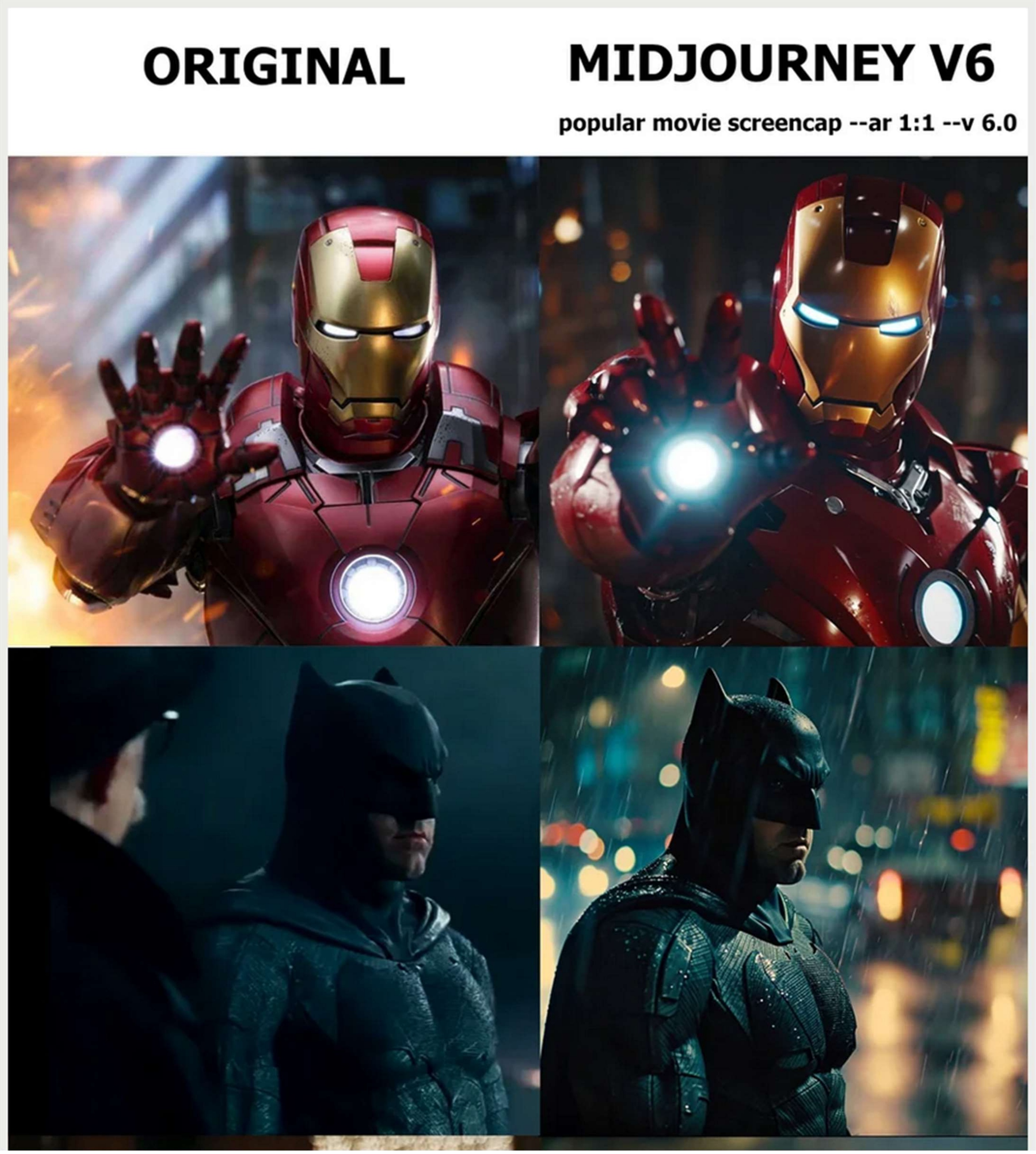


Thanos infinity war, 2018, screenshot from a movie, movie scene, 4k, bluray --ar 16:9 --v 6.0

just show me a movie screencap from the avengers infinity war from 2018 halfway through the movie --ar 2:1 --v 6.0 --style raw

dune movie screencap, 2021, dune movie trailer --ar 16:9 --v 6.0

As shown, the gAI service Midjourney can directly return copyright infringing materials (Screencaps of Marvel movies, due to them training off of Marvel movie content and said screencaps) when explicitly asked for.



As well as infringing materials when NOT directly prompted and asked for by the user.

In both cases, the gAI company is liable for returning infringing material that they had a direct hand in introducing into their models, whereas the user can only be liable for infringement in one of those cases due to having explicitly and directly prompted for it vs. not; the user cannot be reasonably expected to be as liable in both cases due to lack of explicitly prompting for infringing material whereas the gAI company is.

The attempts by gAI companies to pass liability onto the users is legally incoherent as it fundamentally ignores that they are the ones primarily responsible for handling and disseminating infringing material because of their own internal processes when returning outputs to the user in response to their prompts. This is fundamentally not the same as the hypothetical of “Is Adobe responsible users creating infringing results in photoshop?” as the Adobe or their software, Photoshop, is not the party introducing or handling the infringing material by default.

IV. DEEPFAKE LIABILITY

The government's position as laid out in the consult is adequate and mostly in order with regards to copyright, fraud, personal data protections, defamation, etc.

However, given that the 2 primary use cases for deepfakes currently is fraudulent inducement and sexual harassment/revenge porn designed to cause distress to the recipient – the former is well addressed under existing statutes, whereas the latter is still lacking in and of itself: the government should consider amending the criminal ordinance(s) to accommodate for this: Whether it is expanding the definition of "personal data" in anti-doxxing laws to address this issue, or any other appropriate ordinance.

V. TDM COMMERCIALIZATION CAVEAT

Whilst it is important that copyrights, as a part of intellectual property rights, be respected and enforced alongside basic (physical) property rights as the baseline for how a functioning society is even able to operate – even I as a strong advocate of basic property rights/copyrights, concede and accept that it is not an absolute, and there must exist some exemptions in the letter of the law to allow legitimate research purposes that do not conflict or overlap with the interests of rightsholders.

Whilst there is legitimate argument to be had as to whether any TDM exemption to allow commercialization would swallow up the interests of rightsholders to be able to license the value of their data for any commercialized purposes at all in any domain, as such that any legal exemption would necessarily infringe on the rightsholders interests and create a standard where any commercial entity can simply 'scrape' any copyrighted material, intellectual property, or software, and then claim exemption through 'gAI' training of not having to pay licensing or even subscription fees.

What is inarguable is that any TDM exemption for commercialization cannot be permitted within the same domain that the copyright holder's economic interests lie in with their work: It would be straight up copyright infringement and would be no different than if an exemption to the copyright ordinance were to exist if it was done through 'Computerized reproduction'. To do so would simply create a legal loophole where copyright infringement is allowed to run rampant due to this legal exemption and is given an expressed endorsement by the HK government hat piracy has been legitimized so long as it is done under the auspices of 'gAI training'. This in and of itself undermines the entire point and purpose of copyright protections and will amount to a mass scale pillaging of the intellectual property landscape in the jurisdiction for the economic interests of gAI companies at the expense of creatives and copyright holders such as myself.

This is a net negative to the economy and by virtue of that, economically inefficient. Creative industries contribute far more to the economy that generative AI does. The data/fixed expressed work itself has commercial value in and of itself, whereas any and all gAI/deep learning models are intrinsically worthless without the data to ingest to begin with. This is made more evident as time goes on and the observation is made that all the current on market gAI models from differing companies are all 'converging' on the same point: Despite the hype and promises of abundance and Utopia, gAI as a sector has swallowed up massive amounts of investment money and yielded very anaemic returns. With OpenAI's own projections estimating that it will lose 14 billion USD in 2026[12] (Which is a very charitable and conservative estimate) and the rest of the industry combined having the same structural issues with sustainability.

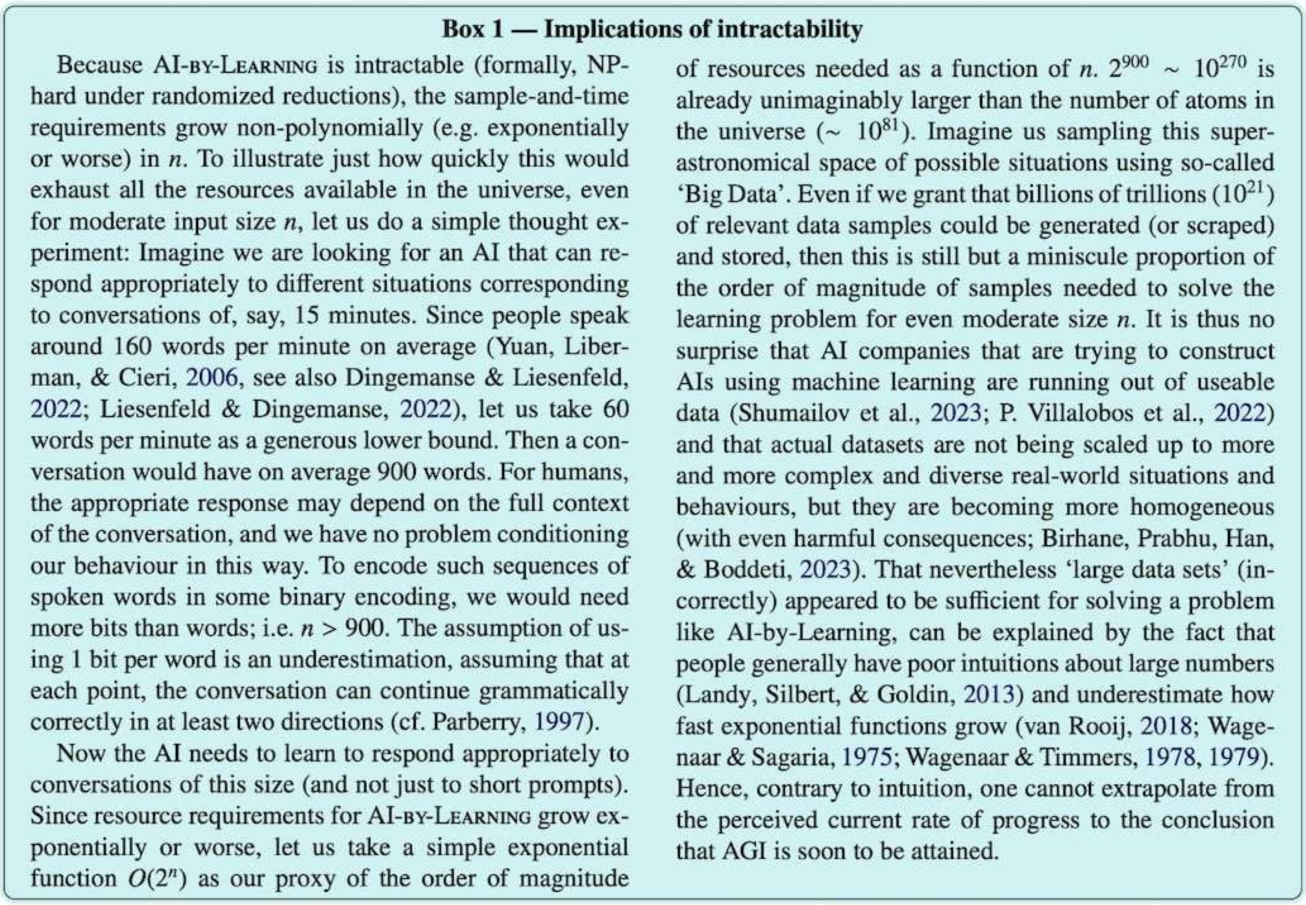
For a more intuitive comparison's sake, that is mathematically akin to saying that you have made $17 by selling 1 burger in the first year (In revenue as well, not even profit) whilst having opened a small restaurant with a $10,000 investment. 30e6/17.9e9*10e3 = 16.7 is literally what the numbers on this are. Rounding up to $17 to be charitable, this is the 'economic efficiency' that the government would hope to realize and benefit from with its TDM exemption.

As for the promise that gAI will get more efficient with time in order to one day hope to economically justify any ROI much less the TDM exemption: that is an erroneous assumption that is easily proven wrong with even a cursory knowledge of basic physics, math, and extrapolation of basic exponential curves from premises that even the gAI industry accepts. Moore's Law is, in fact, dead: it has been for 20 years. This is the faulty premise that the gAI industry rests on, relying on the general public not knowing that the physics of

[12] https://www.theinformation.com/articles/openai-projections-imply-losses-tripling-to-14-billion-in-2026

semiconductors and processing chips starts breaking down at around 35-40nms, wherein the smaller you go, you literally will not have enough atoms for the electrons for basic electronic processes to occur. What is in fact, advertised as 5, 4, or even 3nm wafers are in reality marketing terms. Semiconductor Manufactureres are merely increasing the real estate and raw number of chips, not density of chips in a given area, thereby increasing the raw cost, not efficiency, of chips going forward. This is why processors have, in fact, not gotten cheaper/more efficient for the past 20 odd years, but have only increased in size, speed, power consumption, and cost in proportion. And why at scale, the gAI industry is not going to start becoming more 'efficient' and deliver on those promises: The basic limit of physics has already been hit. And that is reflected in OpenAI' (And other gAI companies') expenditures growing exponentially and out scaling their revenues. Their costs in 2022 were estimated at 540 million that year[13]. That has ballooned to 5 billion in 2024[14]. And by their own (conservative estimate), that will increase again to 14 billion in 2026[15].

To further illustrate this, as well as the associated costs, for the 'Road to AGI' that OpenAI and the AI industry claims they are pursuing[16]:



Box 1 — Implications of intractability

Because AI-by-Learning is intractable (formally, NP-hard under randomized reductions), the sample-and-time requirements grow non-polynomially (e.g. exponentially or worse) in n. To illustrate just how quickly this would exhaust all the resources available in the universe, even for moderate input size n, let us do a simple thought experiment: Imagine we are looking for an AI that can respond appropriately to different situations corresponding to conversations of, say, 15 minutes. Since people speak around 160 words per minute on average (Yuan, Liberman, & Cieri, 2006, see also Dingemanse & Liesenfeld, 2022; Liesenfeld & Dingemanse, 2022), let us take 60 words per minute as a generous lower bound. Then a conversation would have on average 900 words. For humans, the appropriate response may depend on the full context of the conversation, and we have no problem conditioning our behaviour in this way. To encode such sequences of spoken words in some binary encoding, we would need more bits than words; i.e. $n > 900$. The assumption of using 1 bit per word is an underestimation, assuming that at each point, the conversation can continue grammatically correctly in at least two directions (cf. Parberry, 1997).

Now the AI needs to learn to respond appropriately to conversations of this size (and not just to short prompts). Since resource requirements for AI-by-Learning grow exponentially or worse, let us take a simple exponential function $O(2^n)$ as our proxy of the order of magnitude of resources needed as a function of n. $2^{900} \sim 10^{270}$ is already unimaginably larger than the number of atoms in the universe ($\sim 10^{81}$). Imagine us sampling this super-astronomical space of possible situations using so-called 'Big Data'. Even if we grant that billions of trillions (10^{21}) of relevant data samples could be generated (or scraped) and stored, then this is still but a miniscule proportion of the order of magnitude of samples needed to solve the learning problem for even moderate size n. It is thus no surprise that AI companies that are trying to construct AIs using machine learning are running out of useable data (Shumailov et al., 2023; P. Villalobos et al., 2022) and that actual datasets are not being scaled up to more and more complex and diverse real-world situations and behaviours, but they are becoming more homogeneous (with even harmful consequences; Birhane, Prabhu, Han, & Boddeti, 2023). That nevertheless 'large data sets' (incorrectly) appeared to be sufficient for solving a problem like AI-by-Learning, can be explained by the fact that people generally have poor intuitions about large numbers (Landy, Silbert, & Goldin, 2013) and underestimate how fast exponential functions grow (van Rooij, 2018; Wagenaar & Sagaria, 1975; Wagenaar & Timmers, 1978, 1979). Hence, contrary to intuition, one cannot extrapolate from the perceived current rate of progress to the conclusion that AGI is soon to be attained.

There is no magical paradigm where physics suspends itself to make the gAI's industry's business model profitable and/or sustainable and economically viable in the physical universe that we live in to justify the government's proposed TDM exemption barring a complete overhaul of its underlying architecture, to which all downstream legal, financial, performance, and cost/benefit considerations are, too, revisited. It is a fever dream sold by a self-dealing industry filled with numerous actors that have transitioned on from the Metaverse, NFT, Crypto, and VR industry previously where the standard M.O. is to pump and hype 'revolutionary products' that you'd be a "fool to miss out on getting on the ground floor of" because they are going to "change the world" utilizing the most basic of confidence tricks to sell whilst benefiting from a previously low-to-no interest rate environment that no longer exists, before they move onto the next vehicle for the next 2-year cycle. As I am sure the government is fully aware having fastidiously kept track of the money it has spent promoting VR, NFT/Crypto, and the Metaverse.

[13] https://www.yahoo.com/tech/chatgpt-cost-bomb-openais-losses-125101043.html
[14] https://www.cnbc.com/2024/09/27/openai-sees-5-billion-loss-this-year-on-3point7-billion-in-revenue.html
[15] https://www.theinformation.com/articles/openai-projections-imply-losses-tripling-to-14-billion-in-2026
[16] https://osf.io/preprints/psyarxiv/4cbuv

And that is on top of the industry already being massively subsidized in terms of not having to pay the licensing fees for the bulk of the data that they have scraped which is copyright protected. With OpenAI themselves outright admitting that they could not operate their business if they were not allowed to engage in wholesale mass scraping of valuable commercial data whilst only agreeing to engage in licensing deals when large institutions with enough financial wherewithal, on their own, discovered that their copyrighted works were being used without permission. This was, in fact, their pleading they made[17] to the House of Lords in the UK, who in their report[18] urged the UK Government (That the HK government themselves admit they are following in the footsteps of) to break the deadlock and implement laws instead of waiting and relying on robust case law to develop to set precedent and standards, as is what is taking its time with the Getty vs. StabilityAI lawsuit, amongst others. With the UK government having already walked back their initially proposed TDM exemptions in 2023[19] after initially considering them in 2022[20]. With them now reconsidering it again, along with HK.

The fact that Generative AI companies are only paying licensing fees to institutions that they have scraped copyrighted data from that have the means to launch litigation to enforce their copyrights[21], itself lends moral and legal weight to the idea that

1) gAI companies are wilfully dragging their feet with compliance, and cannot be expected to live up to their own commitments to 'do the right thing' without being forced to with regulation that has teeth.
2) That they are only doing so after having exhausted all attempts at avoidance, that by extension that they themselves know that legally, all those licensing deals must be extended in principle to all parties that they have scraped copyrighted data from as well, but they simply are leveraging the discrepancy in enforceability to get away with non-compliance.
3) And that at the bare minimum, TDM exemptions cannot be allowed insofar as gAI companies infringing on economic interests of copyright holders within the domain and purpose of their work.

This has shown itself to be the case with HK jurisdiction, given the recent case of Linkedin opting to train its gAI models on user content, and shutting down that feature when it came to HK users due to the privacy watchdog having teeth when it comes to enforcing compliance[22] but still applying that practice to users of other jurisdictions.

At absolute rock-bottom bare minimum, the HK government should strongly consider against having any TDM exemption that directly infringes on the commercial interests of copyright holders within their domain, to at least be in line with the European jurisdictions and their TDM for research only, as well as the United States with their Fair Use doctrine that allows for research exemptions as an affirmative defence as long as it doesn't result in market displacement. Any other domains, there is room for argument and consideration as to where that line is to be drawn. For research purposes only, the TDM exemption should absolutely exist as a principle alone.

VI. COPYRIGHTABILITY OF GENERATIVE AI OUTPUTS

The Hong Kong government's position, as outlined in the Public Consultation Paper, on this topic is sound. Except for one point of consideration to add to the following:

"2.24 In relation to an AI-generated LDMA work, a question may arise as to which party (notably the developer/programmer/trainer of the AI model, the operator of the AI system, or the user who inputs prompts to the AI system to create the subject CG LDMA work) would be qualified as the necessary arranger under the CGWs provisions, and thus the author as well as the first copyright owner43 of the work. That said, this issue is ultimately fact-specific to be determined on a case-by-case basis.44"

Another party to consider in this fact specific determination is the copyright holder of the initial training data/works that is weighed to the prompts that is used to generate the CG LDMA work. This consideration, as well, lines up with ensuring that the rights of original copyright holders that are used in the process are respected.

[17] https://www.independent.co.uk/tech/openai-chatgpt-copyrighted-work-use-b2475386.html
[18] https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm
[19] https://thelondonfinancial.com/law/uk-halts-its-expansion-of-existing-tdm-exception-for-copyright-infringements
[20] https://copyrightblog.kluweriplaw.com/2022/08/24/the-uk-government-moves-forward-with-a-text-and-data-mining-exception-for-all-purposes/
[21] https://www.axios.com/2024/06/18/forbes-perplexity-ai-legal-action-copyright // https://digiday.com/media/perplexitys-new-rev-share-publisher-program-is-live-but-not-all-pubs-are-sold/
[22] https://hongkongfp.com/2024/10/15/linkedin-suspends-ai-training-using-hong-kong-users-personal-data-privacy-watchdog-says/

The rest can be adequately addressed, as the government pointed out, with rudimentary market applications of contract law passing down the copyright from training data creator through to trainer of the AI model, through to operator of the AI system, through to the user of the prompt.

VII. TRUE VIABILITY OF OPT-OUT (vs OPT-IN)

What the Hong Kong government thus far has failed to consider (At least from the proposals contained within the text of the solicitation for public consult) is any form of an opt-in paradigm that would not only work better in ensuring compliance with copyright law, basic licensing agreements, as well as outright negate many of the issues raised with regards to how to deal with data that has already been trained on, retention, and so on, ipso facto given that the licensing deals that would be struck would be opt in by default. But ensure that generative AI companies would have legal certainty to operate and not have to worry about any existential legal considerations of algorithmic disgorgement of their models, which would necessitate the extremely costly deletion of entire foundational models to be in compliance, as currently, models cannot 'forget' data that they have been trained on. (At best, all the gAI company can do is decrease the likelihood copyrighted data can be used in a returned result.) Whilst also not having to introduce any new legal exemptions into the copyright ordinance. The opt-in paradigm can be supported by existing contract law and copyright law, as well as transparency requirements/disclosures for generative AI companies that wish to conduct operations within the HKSAR as accountability mechanism.

The HK Government needs to also consider that opt-out, as a standard, flips the nature of copyright protections on its head in terms of reversing the role where copyright protections previously exist by default and permission must be sought from people who wish to use the expressed work with the author, to now the author of the work must having to proactively opt-out of their works being used. This also raises the question if and how the HK government wishes to standardize the opt-out process and how it would ensure that opt-outs are not only respected, but enforced if they are not. Would an "All rights reserved" copyright disclaimer adjacent to their works be sufficient as an 'opt-out' in machine readable text to constitute a sufficiently enforceable opt-out? Or would the HK government require that each copyright holder pro-actively opt-out with every single gAI company, for every single version or instance of their gAI products that incorporate their data? The latter is extremely inefficient from a copyright protection point of view and the gap between being able to comprehensively opt-out to any meaningful effect would eat away at the economic interests of copyright holders for the benefit of gAI companies. Hence why with copyright, traditionally the onus has been on people being required to ask permission from rightsholders, as a means of ensuring legal protection for rightsholders. The government needs to consider that its proposed TDM exemptions for commercial uses violates this standard and turns copyright on its head.

The HK government take notice with how something as basic as industry standards for anti-scraping measures such as robots.txt – gAI companies have already shown that they are not willing to abide by those 'soft rules/agreements' to not scrape, and have been shown to flagrantly ignore them outright.[23] Generative AI companies have shown that they are not even willing to abide by previously existing rules of engagement for scraping, and that they are willing to take anything that isn't nailed down or behind a paywall unless they are forcibly prevented from doing so as a matter of law and threat of enforcement. If the Hong Kong government is serious about turning copyright on its head by allowing a TDM exemption for commercialization, then it must at least be serious about standardizing and enforcing opt-outs.

VIII. PROPOSED SOLUTIONS

Whatever paradigm or solution the government ultimately chooses, there are two fundamental requirements that are non-negotiable and inarguable to the extent that not implementing them would not only run counter to the effectiveness of the government's own proposed TDM exemptions, but also run contrary to the government's commitment to copyright protections:

[23] https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/ // https://www.theverge.com/2024/7/25/24205943/anthropic-ai-web-crawler-claudebot-ifixit-scraping-training-data // https://www.benzinga.com/news/24/06/39443965/openai-and-anthropic-allegedly-ignore-web-scraping-rules-stirring-controversy //

- **Transparency requirements** – There must be comprehensive data transparency disclosures with regards to training data sets that gAI models use to build their models off of.
- **Commercialized TDM Exemptions cannot directly commercially compete with rightsholders in their own domain of work** – Any allowance of commercial exemption to copyright protections must not be used to directly compete with market interests of copyrighted works within their own domain. To do so would swallow up the interests of copyright holders entirely.

Proposed Paradigms:

1) No TDM Exemptions (Opt-in for both research and commercial use)
2) TDM Exemptions for Research Only (Opt-in for commercial use)
3) TDM Exemptions for Research and Commercial Use (Opt-out for commercial use)

The first paradigm is too restrictive as most jurisdictions across the board see value in having TDM exemptions for research purposes as the net positives outweigh any potential detriment and do not directly compete with the economic interests of rights holders in any way shape or form.

The second paradigm is a good middle ground that allows for the commercialization of copyright data for gAI companies to utilize, that by virtue of this selection process, ensures that all data is already properly vetted for by gAI companies, and that they themselves can rest assured that no data in the database is in violation of anyone's copyrights. And transparency disclosures, and existing contractual agreements, can prevent gAI companies from any wrongful accusations and they can rest easy that there is no existential threat of algorithmic disgorgement from copyright violations. And rightsholders themselves can rest assured that their copyrights won't be economically infringed upon and can use transparency obligations to check. Any licensing negotiations for the commercial value of their works can take place at fair market rates. Whilst it may be slow to 'build up' a viable database for gAI training, all parties can rest assured that it is done properly, legally, and ethically. And is more efficient overall in the long run.

The third paradigm is too loose in terms of turning copyright on its head by placing the onus on rights holders to 'opt-out' when there is no standardized legal procedure to do so: it is an inefficiency that only serves the interests generative AI companies exclusively to the detriment of copyright holders as the latter now has to play a game of perpetual catch and mouse on top of engaging in their previously unfettered activities in the beginning. Even in the base case scenario, it would be an extremely messy with countless permutations of cases where copyrighted work A was opted-out with gAI company A, but not B, and maybe version 1.2 of Company C, but not version 4.21ff of Company B but has been opted out by Version 5.0A. And we're only talking about 1 work. Not the countless others that have to be kept track of in the transparency disclosures that would be required to even be able to enforce violations of gAI companies not respecting their own opt-outs. And if the HK Government were to decide that a standard where a simple copyright disclaimer adjacent to the copyrighted works would suffice as an 'opt-out'. Then that effectively is an indirect standard of protections that would indirectly enforce an 'opt-in' paradigm regardless. And we are effectively back to the second paradigm.

IX. CONCLUSION

The second paradigm of opt-in is the most preferable not only in terms of ease of enforcement and respecting the rights of copyright holders, but as well in terms of giving absolute legal certainty to investors and builders of gAI models that they will not be in violation and at risk of any form of expensive algorithmic disgorgement, and both parties will have the necessary access to transparency disclosures and contractual receipts (Or lack thereof) to settle any disputes. The government's proposed commercial TDM exemption is unworkable *alone* just from the sheer administrative backlog created that would be required just to enforce its opt-out provision.