

# State of Israel Ministry of Justice

## Legislation and Legal Counsel (Civil Law)

Jerusalem: 25 Kislev 5783

December 18, 2022

## **OPINION: USES OF COPYRIGHTED MATERIALS FOR MACHINE LEARNING**

### **BACKGROUND**

This Opinion originated in a query made by the Israeli Directorate of Defense Research & Development in the Ministry of Defense (DDR&D) to the Ministry of Justice, Legal Counsel and Legislation Department (Civil). The query concerned the flagship Government program for artificial intelligence infrastructures, that is being developed in collaboration with DDR&D, the Israel Innovation Authority, the Higher Education Council, the Ministry of Innovation, Science and Technology and the Ministry of Finance (the “Program”). In this program, DDR&D is responsible for developing Natural Language Processing (NLP) models in Hebrew and in Arabic. This Program was conceived after Israel identified a market failure in the field of NLP in Hebrew and Arabic that stems, *inter alia*, from the insignificant market for Hebrew and Arabic (in the regional dialect) speakers. This small market has reduced market incentives to generate NLP models in these languages and resulted in inferior NLP-based systems in Hebrew and Arabic. To address this problem, the Program took upon itself to absorb the initial high costs of large pre-trained language models and to develop such models itself. Academic, government and private enterprises will then be able to use the models to develop specific NLP use-cases.

The question that was presented to us concerned whether the Program can train its models on copyrighted materials. As the research began, we found that the uncertainty surrounding the ability to use copyrighted materials for machine learning is an acute legal challenge both in Israel and globally. Consequently, we decided to publish an opinion that will address this question broadly.

After writing a first draft, the Counsel and Legislation Department (Civil Law) held an open roundtable (online) that presented the key points in the draft Opinion to Government officials, academics, market actors and members of the public. Following the roundtable, written positions were received from Google Israel, ACUM – Association of Composers, Authors and Publishers of Music in Israel and TALI – The Collecting Society of Film and Television Creators in Israel. Together with oral comments that were received at the roundtable, these responses were taken into consideration and contributed to the formation of this Opinion.

The Opinion was written by Adv. Dr. Lital Helman, under the supervision of Adv. Howard Poliner, Head of the Intellectual Property Department, and the guidance of Adv. Carmit Yulis, the Deputy Attorney General (Civil Law). We wish to thank the Attorney General Adv. Gali Baharav-Miara and the Director-General of the Ministry of Justice Adv. Eran Davidi, who provided valuable support in this effort.

## State of Israel Ministry of Justice

For insightful comments, suggestions and discussions, we are grateful to (in alphabetic order) Adv. Naomi Abraham, Adv. Ofir Alon (Executive Director of the Israel Patent Office), Dr. Roi Baharad, Adv. Zemer Blondheim, Adv. Omri Ben-Zvi, Adv. Eran Bareket, Mr. Elad Dvir, Mr. Eran Dahan, Adv. Ayelet Feldman, Prof. Peter Ficht, Prof. Orit Fischman-Afori, Adv. Yossef Gedaliahu, Ms. Hodaya Gaheli-Schwartz, Adv. Tony Greenman, Adv. Tamar Ganonian-Perkel, Dr. Ziv Katzir, Prof. Moshe Koppel, Dr. Adi Libson, Adv. Noa Mushayeff, Prof. Miriam Markowitz-Bitton, Mr. Barak Peleg, Adv. Guy Paradis, Prof. Gideon Parchomovsky, Adv. Cedric Yehuda Seva, Adv. Haim Ravia, Adv. Dr. Yuval Roitman, Mr. Nir Yanovsky, Mr. Dror Zamir, and Adv. Dr. Efi Zemach, as well as to the participants of the Legislation and Legal Counsel (Civil Law) monthly meeting, the Future of Copyright in the Shadow of New Technologies roundtable in Tel Aviv University chaired by Prof. Amir Houry, the Ono Academic College Faculty Workshop, the IP and Frontier Technologies workshop at the World Intellectual Property Organization, and the Innovation Authority's meetup, chaired by Dr. Ziv Katzir, head of the international program on artificial intelligence infrastructures. We wish to thank also Adv. Meir Levin, the Deputy of the Attorney General (Financial), who directs the work on artificial intelligence in the Ministry of Justice, and Adv. Zafir Neumann, the legal counsel of the Israel Innovation Authority. A special gratitude is extended to Tal Geva from DDR&D for his professional advice in the field of machine learning.

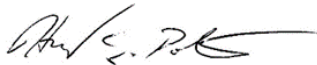
It is our hope that this Opinion will be followed in case law and will serve to enhance the development of both machine learning and creativity and culture in Israel.



---

Adv. Carmit Yulis

Deputy Attorney  
General (Civil Law)



---

Adv. Howard Poliner

Head of the Intellectual  
Property Department



---

Dr. Adv. Lital Helman

Intellectual Property  
Department

# State of Israel Ministry of Justice

## ABSTRACT

This Opinion aims to shed light on the most fundamental question in the intersection between machine learning (ML) and copyright law: whether ML enterprises can make unauthorized use of copyrighted materials to train Artificial Intelligence (AI) systems. ML is the process that enables computers to autonomously learn from past data. ML thus provides the foundations for AI systems. The intersection between ML and copyright bears crucial importance. ML is becoming increasingly central to the global economy and to Israel in particular, and Israel holds a leading position as a producer of AI systems. Lifting copyright uncertainties that surround this issue can spur innovation and maximize the competitiveness of Israeli-based enterprises in both ML and content creation.

As this Opinion explicates, the value of AI systems depends first and foremost on the quantity of the materials that the machine receives at the ML stage, together with the diversity and quality of such materials. ML enterprises may use materials of various kinds—photos, text, sound, video, etc.—depending on the task that the machine learns to eventually perform.

Obviously, copyright law imposes no limitation on training machines on public domain materials or on works that are copyrighted or licensed by the ML enterprise that uses them. But given the immense quantity of works that ML utilizes, the ML process will almost always require the use of works that are copyrighted by multitudes of third parties. Current Israeli copyright law provides little guidance as to the infringing status of such a use. The vagueness of copyright law on this crucial issue yields a severe legal uncertainty that may hinder the growth of ML and AI, without yielding substantial benefits for copyright owners. In fact, this legal uncertainty may also form a hurdle for copyright enforcement in this arena.

This Opinion concludes that apart from certain circumstances, the use of copyrighted materials for ML is permitted under existing copyright doctrines. First and foremost, ML will typically be covered by the fair use doctrine. Second, some ML projects may fall under the doctrine that permits incidental uses of copyrighted materials. Third, in the (admittedly nonstandard) case where the copyrighted materials are erased at the end of the ML process, the enterprise may enjoy the protection provided under the doctrine of transient use. Notably, the conclusion of this Opinion—that training ML systems is generally permitted under copyright law—is consistent with the approach of other legal systems around the globe.

While most ML cases come within the scope of permitted uses, this Opinion specifically excludes from its scope certain ML uses. For example, the safe harbor set forth in this Opinion would not apply to ML datasets that consist exclusively of works created by a single author in order to compete with this author in her existing markets. As a general matter, this Opinion does not apply to the output of the ML process. There may indeed be cases where the ML process would be protected under this Opinion, yet the output (or some of the outputs) of the resulting AI system would be infringing. Overall, this Opinion endeavors to strike the right balance that will stimulate innovation in the ML domain while at the same time maintaining copyright integrity and enhance the incentive to create new works.

# State of Israel Ministry of Justice

## OPINION: USES OF COPYRIGHTED MATERIALS FOR MACHINE LEARNING

### A. INTRODUCTION

Machine learning (ML) is one of the most dramatic revolutions of the 21<sup>st</sup> century. This revolution promises significant changes in all parts of society and economy, ranging from labor markets, transportation and real estate to art, defense, and medicine and to all aspects of society.<sup>1</sup> ML is expected to contribute approximately 15.7 trillion USD to the global economy by 2030, and countries around the world compete for their share of this enormous projected growth.<sup>2</sup> The State of Israel is at the forefront of ML development, owing to local start-ups, domestic and foreign investments, and cutting-edge Research and Development centers global companies have established in the country.<sup>3</sup>

Naturally, the novelty that ML brings, leaves various legal questions unanswered under the extant law.<sup>4</sup> This Opinion tackles one such question, and perhaps the most fundamental one. As explained further below, artificial intelligence (AI) systems are based on enormous datasets as the basis for learning. The problem is that these datasets typically

---

<sup>1</sup> See, e.g., Avraham Tenenbaum, *Fundamental Guidelines for Law Governing Autonomous Vehicles*, MISHPAT MAFTTEACH, 3, 33, 36 (2015) (in Hebrew); Winnie Hu, *Driverless Cars Arrive in New York City*, N.Y. TIMES (Aug. 6, 2019), <https://www.nytimes.com/2019/08/06/nyregion/driverless-cars-new-york-city.html> (traffic sector); Nicola Davis, *AI Equal with Human Experts in Medical Diagnosis, Study Finds*, GUARDIAN (Sept. 24, 2019), <https://www.theguardian.com/technology/2019/sep/24/ai-equal-with-human-experts-in-medical-diagnosis-study-finds> (medical sector); Jason Pontin, *How AI-driven Insurance Could Reduce Gun Violence*, WIRED (Feb. 27, 2018), <https://www.wired.com/story/how-ai-driven-insurance-could-reduce-gun-violence> (insurance sector); Fortune Business Insights, *Home Automation Market to Expand at 12.1% CAGR and Reach USD 114 Billion by 2025*, GLOBENEWSWIRE (Apr. 27, 2021), <https://www.globenewswire.com/news-release/2021/04/27/2217431/0/en/Home-Automation-Market-to-Expand-at-12-1-CAGR-and-Rreach-USD-114-Billion-by-2025.html> (real estate sector); Amir Mizroch, *In Israel, a Standout Year for Artificial Intelligence Technologies*, FORBES, (Mar. 11, 2019), <https://www.forbes.com/sites/startupnationcentral/2019/03/11/in-israel-a-stand-out-year-for-artificial-intelligence-technologies/?sh=548db8ba30a8> (defense sector).

<sup>2</sup> See Global Artificial Intelligence Study: Exploiting the AI Revolution' Sizing the prize: PwC Global, <https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html>; Neil Savage, *The race to the top among the world's leaders in artificial intelligence*, NATURE INDEX, Sept. 9, 2020, <https://www.nature.com/articles/d41586-020-03409-8>.

<sup>3</sup> See Liran Antebi, *Artificial Intelligence and National Security in Israel* (INSS 2020) 85; DataNation, *Record year: AI start-ups raised 25.2 billion dollars in 2018*, THE MARKER, 3.21.19, <https://www.themarker.com/technation/datanation/1.7017976> (in Hebrew).

<sup>4</sup> See CA 9183/09 *The Football Association Premier League Limited v. John Doe*, SCR 65(3) 521, para. 5 in the opinion of J. Meltzer (2012) ("It is known that usually technology is way ahead of the law. In such circumstances, the legislator and courts are required to pour the existing, good and substantiated principles into new legal vessels, as if it is wine that gets better with time and needs only a vessel") (hereinafter: the "Premier League Case"); APA 3782/12 *Commander of the Tel Aviv - Yafo District in Israel Police v. Israel Internet Association* (published in Nevo, 3.2.2013) para. 23 in the opinion of J. Solberg (2013) ("As known, the law lags behind innovations, and legislation does not meet the pace of progress of science and its applications.").

## State of Israel Ministry of Justice

contain content that was created by others and is therefore copyrighted. Therefore, the underlying question of this Opinion is one of the basic questions pertaining to ML: to what degree does copyright allow AI to learn?

The general rule of copyright law is that the reproduction and other specified uses of works require the permission of the works' copyright owners.<sup>5</sup> This general rule has exceptions. First, a work may be in the public domain, either because its copyright has expired or because no copyright subsisted in the work in the first place.<sup>6</sup> Second, a work may be distributed under a Creative Commons (CC) license, which enables the use of the work under the terms set forth in the license.<sup>7</sup> Third, and most importantly for our purposes, copyright law *itself* permits certain uses of copyrighted works with neither permission of nor payment to rightsowners.<sup>8</sup> Indeed, copyright law deems absolute protection of works undesirable.<sup>9</sup> Copyright law aspires to strike a balance between the interests of authors in restricting access to works and the public interest (and the interest of future authors) in increasing access to works.<sup>10</sup> For this reason, copyright is inherently limited and restricted, and comprises both exclusive rights in works and limitations on these rights.<sup>11</sup>

---

<sup>5</sup> The copyright owner is the author, the person the author transferred the copyright to, or whoever the owner of the copyright is by operation of the law. *See* §§ 11, 33-37 of the Copyright Act 2007.

<sup>6</sup> *See* §§ 4-5 of the Copyright Act. *See also* Proposal for a DIRECTIVE of the EUROPEAN PARLIAMENT and of the COUNCIL on copyright in the Digital Single Market (2016) 593 final, 2016/0280, 14 September 2016, COM (Recital 8) (“[t]ext and data mining may also be carried out in relation to mere facts or data which are not protected by copyright and in such instances no authorization would be required”).

<sup>7</sup> CC licenses change the default from “all rights reserved” to “some rights reserved” and permit the use of works under conditions such as credit to the author (CC-BY); limitation on commercial use (CC-NC) or limitation on a derivative work (in general or under a ‘share alike’ obligation). For CC licenses see Creative Commons, <https://creativecommons.org/about/cclicenses/>.

<sup>8</sup> *See* §§ 19-30 of the Copyright Act; § 33 of the Copyright Ordinance, 1911. *See also infra* Part C.2.

<sup>9</sup> *See, e.g.,* CA 8393/96 *Mifal HaPaysis v. The Roy Export Establishment Company*, SCR 54(1) 577, 596 (2000) (“when coming to protect the original work it is necessary to take into consideration that a protection that is excessive might curtail cultural and social development that rests on past achievements.”) (hereinafter: the “Mifal HaPaysis Case”). *See also* Gideon Parchomovsky & Abraham Bell, *Reinventing Copyright and Patent*, MICH. L. REV., Vol. 113, P. 231 (2014).

<sup>10</sup> *See* explanatory notes to the Copyright Act, General Part, 5768-2007, Government Bill 1116 (“Copyright laws aims to protect works, while striking a balance between different interests in favor of the public”); 5097/11 *Telran Communications (1986) Ltd. v. Charlton Ltd.*, para. 13 in the judgment of J. Zylbertal (Nevo, 2.9.2013) (defining copyright as the product of “balances between the interests of future authors and of users in a ‘creative maneuvering space’”) (hereinafter: the “Telran Case”); CA 326/00 *Holon Municipality v. NMC Music Ltd.*, SCR 57(3) 658, 663-664 (2003) (“and it is here – in the interpretive sphere, that we see the tension in copyright laws between different interests, including, primarily, the property interest of the rightsholder, and the interest of the public”). *See also* Orit Fischman-Afori, *Cultural rights and human rights: a tool for a balanced development of copyright laws in Israel*, 37 MISHPATIM 499 (2007).

<sup>11</sup> Copyright exceptions include, *inter alia*, limitation of the period of copyrights, no-application of copyright to ideas, methods and data, and a list of permitted uses in work with neither license nor payment for the use. *See, e.g.* § 5 and chapter D of the Copyright Act. *See also* TONY GREENMAN, *COPYRIGHT*, (2nd Ed, Vol. 4) (2008) (hereinafter: “Greenman”); Michael Birnhack, *The Legal Work: Fair Use in Copyright Laws*, in *LAW, CULTURE AND BOOK – NILI COHEN BOOK* (O. Grosskopf and S. Lavi Eds. 2017) 1, 1 (“copyright laws aspire to encourage the work by the creation of a legal tool that is intended to facilitate autonomous action in the free market, a kind of a property right, however that is shaped in advance, delimited and limited.”).

## State of Israel Ministry of Justice

The applicability of copyright exceptions to ML has not been clarified in Israel, to the detriment of both the AI and content industries. Uncertainty can form a barrier to the development of high-quality datasets for ML and impose legal costs on AI-based enterprises.<sup>12</sup> Likewise, legal uncertainty might impede enforcement of copyrights in adequate circumstances. This Opinion aims to clarify the law on this issue.

In a nutshell, this Opinion maintains that three copyright limitations potentially apply to ML datasets. First and foremost, the fair use doctrine, set forth in section 19 of the Copyright Act (the *Act* or the *Statute*), typically permits the creation of ML datasets. This interpretation stems from both the language of the section and its purpose and is also consistent with U.S. Law, from which the Israeli law adopted the fair use doctrine.<sup>13</sup> Second, section 22 of the Statute, which discusses incidental works, can apply because works in datasets serve as a means to train the computer to perform a task, and it is the task that is the purpose of the training. This doctrine will not apply in cases where the dataset has a substantial economic value and in cases that exceed the scope of section 22.<sup>14</sup> Third, circumstances (that are not frequent) in which the dataset is deleted at the end of the training process can trigger section 26, which permits transient copying. Although this is an admittedly broad interpretation of the term transient copying, it is consistent with the language of the Statute and with the interpretation of this term under European law.<sup>15</sup>

In conclusion, this Opinion posits that copyright law permits, in most cases, the inclusion of copyrighted materials in ML datasets. This approach meets the purpose of copyright law for two reasons. The first reason lies in the analogy to inductive human learning. Indeed, when a human learns, the dataset that enables learning is ‘located’ in her brain, without interfering with copyright law at all.<sup>16</sup> A computer, however, cannot (at present) learn by ‘reading’ content, unless such content is *copied* first to a dataset that the computer can read.<sup>17</sup> In other words, copyright arises in the context of ML only as a result of a technical-technological limitation of computer learning (that might change as technology advances). But the promise hidden in AI—allowing machines to track human learning and eventually perform equivalent tasks at a high level—can only be fulfilled if learning will be enabled for machines, even though technically the process by which a

---

<sup>12</sup> This concern is not unique to Israel. *See, e.g.*, European Commission (2016), Commission Staff Working Document, Impact Assessment on the modernisation of EU copyright rules, 14 September 2016, SWD (2016) 301 final, Part 3/3, p. 94 (“researchers are faced with legal uncertainty with regard to whether and under which conditions they can carry out TDM on content they have lawful access to.”).

<sup>13</sup> *See* 17 U.S.C. § 107. *See also infra* Part D. For a comparison between fair use in U.S. Law and in Israeli law, see Neil Netanel, *Israeli fair use from an American point of view*, in *CREATING RIGHTS: READINGS IN THE COPYRIGHT ACT 377* (Michael Birnhack and Guy Pesach Eds. 2009).

<sup>14</sup> *See infra* Part C.2.B.

<sup>15</sup> *See infra* Part C.2.C.

<sup>16</sup> *See, e.g.*, Thomas Margoni, *Artificial Intelligence, ML and EU Copyright Statute: Who Owns AI?*, CREATE Working Paper (Dec. 2018), available at SSRN: <https://ssrn.com/abstract=3299523> (“when humans learn a new language they usually store the training information (e.g. the text book used to learn it) as an electrochemical trace in the area of the brain dedicated to language. Humans do not need a copyright exception in order to store that copy.”) (hereinafter: “Margoni”).

<sup>17</sup> *Id.*, at 4 (“In NLP, as well as in most text analytic fields, algorithms “learn” abstract probabilistic models from texts annotated with labels... in order to predict such labels on unseen text. They do this by storing the relevant information in a separate file, the “trained model.”)

## State of Israel Ministry of Justice

machine learns necessitates the making of copies. Granted, a direct analogy between computer and human learning is imperfect. Computer learning is not intended to expand the horizons of human knowledge, but rather to further the financial interests (usually) of the firm that is behind the ML systems. Yet, we believe that learning *per se* should be protected, whether the learner is a human or a machine, because in both scenarios the learning forms the basis for the development of new capabilities and encourages progress. In circumstances where copyright law justifies restriction of certain learning products, such restriction should be imposed at the *output* stage and not at the learning stage, which can also serve as basis for generating legitimate products.

Second, applying copyright limitations to most cases of ML strikes a good balance between the interests of the authors and those of the public, which is the keystone of copyright policy.<sup>18</sup> On the one hand, the application of copyright limitations to ML will allow AI to develop and flourish. Absent copyright limitations, market failures in the field of digital access to works might frustrate the creation of effective datasets. Thus, and as discussed below, transaction costs involved in locating and obtaining licenses from numerous copyright owners, as well as the power of each rightsholder to delay the project (the ‘holdup problem’) can place severe chilling effects on innovation in the field of ML and stifle the operation of AI enterprises.<sup>19</sup>

On the other hand, and as discussed further below, the harm to copyright owners from applying copyright limitations to ML is negligible, if at all. ML datasets do not cause any harm whatsoever to the existing markets of copyright owners, and in most cases, do not deny them of substantial potential profits.<sup>20</sup> Notably, each individual work is a single component in an enormous dataset and holds an immaterial weight in the dataset. The market value of each work will undoubtedly reflect this immateriality and be infinitesimal—probably significantly lower than the costs required for obtaining licenses for each work.

The need to shape proprietary structures dynamically for the purpose of attaining social goals is firmly established.<sup>21</sup> This need is intensified in the context of copyright, because despite copyright’s proprietary characteristics, copyright is different from tangible

---

<sup>18</sup> See *supra* notes 9-10. See also Universal Declaration of Human Rights, GA Res 217 (III) A, UN Doc A/RES/217(III) art 27 (10 Dec. 1948) (recognizing in the first paragraph the right of access to information and in the second paragraph the right of the author to limit access to his works; JAMES BOYLE, SHAMANS, SOFTWARE & SPLEENS: LAW AND THE CONSTRUCTION OF THE INFORMATION SOCIETY 38 (1996); ROBERT P. MERGES, JUSTIFYING INTELLECTUAL PROPERTY 136 (2011).

<sup>19</sup> See *infra* notes 87-91 and accompanying text.

<sup>20</sup> For further information see *infra* Part C.2.a.

<sup>21</sup> See, e.g., LCA 6339/97 *Rocker v. Salomon*, SCR 55(1) 199, 280 (“property is intended to express the control of a person and his personality. Nevertheless, property has a public aspect...and the use of property should serve the public interest”); Tel Aviv District Court 2177-05 *ADIDAS-SALMON v. Yassin* (published in Nevo, 13.12.2010), p. 30 (“it is necessary to shape property rights in a manner that will express the proper social and cultural development”); Hanoch Dagan, *Property reading: the renewing property institution of copyright*, in CREATING RIGHTS – READINGS IN THE COPYRIGHT ACT 47 (2008) (“discourse on property is, by its very own nature, purposive and dynamic: it enables, and maybe even invites – constant development (albeit cautious) of the different property institutions in a manner that will improve their actions in promoting these purposes”).

## State of Israel Ministry of Justice

property and its scope is restricted by inherent limitations and exceptions.<sup>22</sup> The analysis provided in this Opinion, namely that copyright does not ordinarily extend to the creation of ML datasets, reflects adequately, in our opinion, the inherent balances underlying copyright law.

Importantly, this Opinion posits that ML datasets are protected in *most* cases, but not in *all* of them. In particular, the Opinion shields from liability datasets that include vast and diverse works, where each work occupies a negligible and immaterial part. Datasets that purposely comprise of a specific type of works (typically for the purpose of producing identical products) might be excluded from the Opinion, as further elaborated below.

The scope of this Opinion and its limitations are discussed in detail below.<sup>23</sup> Notably, this Opinion does not apply to ML-based *products*, but only to the learning process itself. The infringing status of *the product* will be examined ad-hoc based on extant copyright rules and standards, and this Opinion does not grant products an a-priori safe harbor.<sup>24</sup> It should also be noted that this Opinion does not cover the use of content under other laws, such as privacy laws, immunities, national security, ethics, and the like.<sup>25</sup> Finally, this Opinion covers Israeli law only, and is not intended to apply outside the territory of the State of Israel or when the applicable law is not Israeli law.<sup>26</sup>

The Opinion will unfold as follows: Part B provides a technological background for ML datasets. Part C delineates the normative framework of Israeli copyright law and applies it to the issue at hand. Part D discusses comparative law on the subject. Part E discusses the scope of the Opinion and the cases that lie beyond its scope. Part F concludes the discussion and discusses, *inter alia*, the option to pursue legislative amendments in the tenor of this Opinion. The Opinion concludes that it is preferable to rely on copyright limitations that are laid down in the extant law, at least for the time being.

### B. BACKGROUND: ML AND AI-BASED SYSTEMS

ML is the technological foundation of AI systems.<sup>27</sup> ML is the process by which computers inductively learn from datasets by themselves. As opposed to ‘non-learning’

---

<sup>22</sup> See *supra* note 11.

<sup>23</sup> See *supra* Part D.

<sup>24</sup> See *supra* Part D.

<sup>25</sup> See, e.g., Orin Kerr, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143, 1165 (2016).

<sup>26</sup> The choice of laws rules in copyright laws are complex and exceed the scope of this Opinion. See, e.g., CA 2790-93 *Eisenman v. Kimron*, SCR 54(3) 817 (2000).

<sup>27</sup> The term ML was coined already in 1959. See Arthur L. Samuel, *Some Studies in ML Using the Game of Checkers I.*, in COMPUTER GAMES I 335 (1988). Yet the technological leap that facilitated current ML is dated to approximately 2010. See, e.g., Michael Veale & Irina Brass, *Administration by Algorithm? Public Management Meets Public Sector ML*, in ALGORITHMIC REGULATION 121, 125 (2019) (“a machine learns when its performance at a certain task improves with experience.”).



## State of Israel Ministry of Justice

systems, in which programmers input the data that is required for the software operation in person, in AI-based systems the human involvement is limited to writing the algorithms that guide the computer learning and to the creation of datasets that will serve as the basis for learning. After the computer ‘trains’ on the dataset, it can implement its conclusions to cases that were not presented to it previously.<sup>28</sup> These new cases can later be included in the dataset and become a basis for learning too. Consequently, the computer’s performance consistently improves.<sup>29</sup>

Training a learning system requires large, relevant, and diverse datasets.<sup>30</sup> The data in the dataset may include various types of data—photos, texts, audio files, multimedia files, etc.—depending on the nature of the task that the system learns to perform. Datasets are created in two stages. The first stage includes the collection of data—ordinarily by automatic software (‘bots’). At the second stage, irrelevant information that was collected together with the data, such as design, links etc., is ‘cleansed’.<sup>31</sup> These two stages require the investment of a considerable amount of time and resources and consume the lion’s share of the inputs invested in ML enterprises.<sup>32</sup>

Following these stages, learning algorithms that operate on the data create a statistical model, that can either stand by itself, or serve as a ‘basic layer’ for the development of specific tasks. Thus, for example, a Natural Language Processing (NLP) system will train on a large dataset of textual content. At the end of the training process, the system will create a model that will be able to ‘understand’ natural language in a general manner. Specific tasks will be able to use this model for the purpose of creating various NLP-based systems, such as a system for automated summaries of documents, voice-identification systems, an automatic customer service, and many others.

Specific tasks entail an additional stage—‘labeling’ the dataset, i.e., providing examples that will teach the computer how to learn from the data. For example, a system that is intended to identify animals will be able to train on labeled dataset that is organized as a two-column table: the first column includes the photos, and the second—an explication of the animal displayed in the photo. The system will learn from the labeled dataset to distinguish between animals, and at the end of the training process will be able to identify

---

<sup>28</sup> See Will Knight, *This AI-Generated Musak Shows Us the Limit of Artificial Creativity*, MIT, <https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart>, TECHNOLOGY REVIEW (April 26, 2019); Ayush Pant, INTRODUCTION TO ML; BEGINNERS, TOWARD DATA SCIENCE (Jan. 7, 2019) <https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08>.

<sup>29</sup> *Id.*

<sup>30</sup> See Google, Training and Test Sets: Splitting Data, <https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data>.

<sup>31</sup> See, e.g., JIAWEI HAN, MICHELINE KAMBER & JIAN PEI, DATA MINING: CONCEPT AND TECHNIQUES, 3RD. ED. (2011).

<sup>32</sup> See, e.g., Israel Innovation Authority – Call for Proposals – Creating datasets in spoken Hebrew and/or Arabic [https://innovationisrael.org.il/kol-kore/Infrastructure\\_datasets](https://innovationisrael.org.il/kol-kore/Infrastructure_datasets) (last viewed: April 8, 2022) (A Call for proposals to create a dataset for the NLP enterprise, where the prime budget is granted for collecting and “cleaning” the dataset—ILS 550,000 for each application).

## State of Israel Ministry of Justice

animals in new photos that it did not ‘see’ before. ‘Labeling’ is typically performed by humans or by a combination of human and machine.<sup>33</sup>

Interestingly, a dataset is not a static object, but rather a dynamic work that evolves during the lifetime of the enterprise.<sup>34</sup> Thus, datasets are updated by the addition or removal of data for various purposes, such as enhancements, error correction, legal or other demands to remove items, changes in the system’s use-cases, etc. The dataset is indeed a cardinal and fundamental element of AI systems throughout the lifecycle of the enterprise.<sup>35</sup>

The importance of a good dataset cannot be overemphasized. The quality of AI systems depends first and foremost on the scope of the data the system collects in the ML stage, as well as the quality and diversity of such data.<sup>36</sup> From the prism of innovation, it is crucial to facilitate datasets that are as broad and relevant as possible. On the other hand, copyrights limit the use of materials created by others.<sup>37</sup> Does copyright law inhibit the inclusion of copyrighted materials in datasets for ML? The next Part addresses this question.

---

<sup>33</sup> See Margoni, *supra* note 16, para. II. Learning from unlabeled data is known as unsupervised learning.

<sup>34</sup> See, e.g., MIMIC- III, <https://physionet.org/content/mimiciii/1.4> (last viewed: 5.2.22) (describes a huge project for a medical dataset that is maintained and updated frequently by research and industry entities).

<sup>35</sup> See also Amanda Levendowski, *How Copyright Statute Can Fix Artificial Intelligence’s Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018) (discussing copyright in the datasets themselves, an issue that lies beyond the scope of this Opinion).

<sup>36</sup> The industry dubs this phenomenon ‘Garbage in garbage out.’ See Garbage in garbage out, WORLD WIDE WORDS, <http://www.worldwidewords.org/qa/qa-gar1.htm>. See also studies that demonstrate acute biases in the operation of AI systems based on limited datasets or datasets that are not sufficiently diversified: SARA WACHTER-BOETTCHER, *TECHNICALLY WRONG: SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH* (2017); Wenying Wu et al., *Gender Classification and Bias Mitigation in Facial Images*, in 12th ACM CON. WEB SCIENCE 106 (2020), <https://doi.org/10.1145/3394231.3397900>; Levendowski, *id.*

<sup>37</sup> Protection of works created by nonhuman authors has been debated around the world over the years. See, e.g., *Naruto v. Slater*, No. 16-15469 (regarding a photo that was taken by a chimpanzee). In one of the latest developments in this field, the Library of Congress refused to recognize copyright in a work that was created by a computer. See, Copyright Office, *Second Request for Reconsideration for Refusal to Register a Recent Entrance to Paradise* (Correspondence ID 1-3ZPC6C3; SR # 1-7100387071), Feb. 14, 2022, <https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf>. Courts in China ruled in favor of granting such a protection as said. See Paul Sawers, *Chinese Court Rules AI-Written Article is Protected by Copyright*, VENTUREBEAT (Jan. 10, 2020), <https://venturebeat.com/2020/01/10/chinese-court-rules-ai-written-article-is-protected-by-copyright/>. For literature on the subject see, for example, Pamela Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT. L. REV. 1185 (1986); Clark D. Asay, *Independent Creation in a World of AI*, 14 FLA. INT’L. U. L. REV. 201 (2020); Annemarie Bridy, *Coding Creativity: Copyright and the Artificially Intelligent Author*, 2012 STAN. TECH. L. REV. 5, 2; Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343 (2019); James Grimmelmann, *There’s No Such Thing as a Computer-Authored Work—And It’s a Good Thing, Too*, 39 COLUM. J. L. & ARTS 403 (2016); Liubov Maidanyk, *Artificial Intelligence and Sui Generis Right: A Perspective for Copyright of Ukraine?*, 3(11) ACCESS TO JUSTICE IN EASTERN EUROPE 144–154 (2021).

# State of Israel

## Ministry of Justice

### C. THE NORMATIVE FRAMEWORK – THE USE OF COPYRIGHTED MATERIALS

This Part explicates the normative framework for the use of copyrighted works. As elaborated below, under Israeli law, works are usually copyrighted during the life of their author and seventy years after her death.<sup>38</sup> Copyright confers upon authors exclusive rights to make numerated uses of the work. This means that such uses require permission of the copyright owner. Alongside the exclusive rights regime, copyright law permits uses of works without permission nor payment of consideration for the use in certain cases.<sup>39</sup> The first Section of this Part examines what uses of ML datasets may come under copyright. The second Section discusses what copyright limitations may apply to ML.

#### 1. COPYRIGHT PROTECTION

The Copyright Act defines the scope of copyright protection and incorporates Israel's international obligations. Under the Copyright act, the threshold for copyright protection of works is rather low, namely, copyright is extended to works without serious conditions. Section 4(a) of the Statute states the following –

**“(a) Copyright shall subsist in the following works:  
(1) Original works which are literary works, artistic works, dramatic works, or musical works, fixed in any form;  
(2) Sound recordings;  
provided that the aforesaid works fulfill one of the conditions set forth in section 8 or that copyright subsists in said works pursuant to Order in accordance with section 9.”**

Without delving into the specifics of the Statute, the low threshold of copyright protection means that the collection of extensive and vast amounts of materials by ML enterprises will most likely include copyrighted materials. Together with the fact that copyright lasts seventy years after the death of the author, the low protection threshold explains why virtually all datasets that will be created for ML will contain copyrighted materials. Consider also that copyrighted works are not necessarily labeled as such and are not registered.<sup>40</sup> Therefore, potential users of the works cannot easily discern between

---

<sup>38</sup> Special provisions for copyright terms apply to special cases, such as sound recordings, fonts and works created by the State. *See* §§ 38-44 of the Copyright Act.

<sup>39</sup> *See* chapter D of the Copyright Act.

<sup>40</sup> *See* Article 5(2) of the Berne Convention for the Protection of Literary and Artistic Works, U 21, 581 (opened for signing in 1949) (prohibits the conditioning of copyright protection on formalities, such as

## State of Israel Ministry of Justice

copyrighted and non-copyrighted works when collecting materials. For these reasons, AI enterprises must reasonably assume that at least part of the works that they use in their datasets, and probably the big majority of them, are protected by copyrights.

Copyrighted works are protected by two types of rights: copyrights, pursuant to section 11 of the Statute (also dubbed “economic rights”) and a moral right, pursuant to section 46 of the Statute. Let us now delve into these two types of arrangements.

### A. Copyright

Copyright owners have the exclusive right to make specified uses of the works as defined in the Statute (the “bundle of rights”). Using one or more of the exclusive rights without the copyright holder’s consent is infringing, provided that no relevant exception applies. As Section 11 provides—

**“Copyright in a work means the exclusive right to do with the work, or a substantial part thereof, one or more of the following acts, in accordance with the kind of the work:**

- (1) Reproduction as stated in section 12 – with respect to all categories of works;**
- (2) Publication – in respect of a work not yet published;**
- (3) Public performance as stated in section 13 – in respect of a literary work, dramatic work, musical work and sound recording;**
- (4) Broadcasting as stated in section 14 – in respect of all kinds of works;**
- (5) Making a work available to the public as stated in section 15 – in respect of all kinds of works;**
- (6) Making of a derivative work as stated in section 16 and the doing of any acts set forth in sections (1) to (5) above in respect of the aforesaid derivative work – with respect to a literary work, artistic work, dramatic work and musical work;**
- (7) Rental as stated in section 17 – in respect of a sound recording, cinematographic work and computer program.”**

---

registration). In the United States a registration on copyright exists, but the registration is not a condition for protection. *See* 17 U.S.C. § 408(a) (2018) (“[R]egistration is not a condition of copyright protection.”); *Kernel Records Oy v. Mosley*, 694 F.3d 1294, 1301 (11th Cir. 2012); *Automation by Design, Inc. v. Raybestos Prods. Co.*, 463 F.3d 749, 752 n.1 (7th Cir. 2006).

## State of Israel Ministry of Justice

Technically, the creation of a dataset reproduces works during their collection (a right granted under section 11(1)), stores them in the dataset (technically, another reproduction) and often performs changes in the work. Such changes may be minor or substantial, as ML enterprises sometimes make deliberate manipulations to works for the purpose of creating more than one variation of the work, thus expanding the dataset and its diversity. Making changes in the work might come under the right to make derivative works, per section 11(6). As a result, unless copyright limitations apply, ML dataset are likely to expose the enterprise that operates them to copyright infringement claims *from each of the rightsowners in the many works that are included in the dataset.*

### *B. Moral Rights*

Alongside the economic rights, the Statute confers upon authors moral rights, as set forth in section 46 of the Copyright Act, as follows:

**“A moral right in relation to a work is the right of its author –**

**(1) To have his name identified with his work to the extent and in the manner suitable in the circumstances;**  
**(2) That no distortion shall be made of his work, nor mutilation or other modification, or any other derogatory act in relation to the work, where any aforesaid act would be prejudicial to his honor or reputation.”**

Because datasets do not ordinarily attribute works to authors and because datasets often modify the work, ML enterprises are theoretically also vulnerable to moral rights claim of both the right of attribution under section 46(1) and the right of integrity under section 46(2). This is practically complicated, because moral rights cannot be transferred, and are held by *the author* even if the copyrights in the work are transferred to a third party.<sup>41</sup> For ML enterprises, this means double the risk: legal exposure persists from both copyright owners for copyright infringement and authors for violation of the moral right.

---

<sup>41</sup> See § 45 of the Copyright Act.

## State of Israel Ministry of Justice

Yet moral rights are not absolute.<sup>42</sup> First, they do not apply to all types of works.<sup>43</sup> Second, the right of attribution depends “on the proper scope and degree under the circumstances” and the right of integrity is only violated if the modification of the work is “prejudicial to the author's honor or reputation.”<sup>44</sup> Granted, courts favor a broad interpretation of moral rights, in particular with respect to the right of attribution.<sup>45</sup> However, we are of the opinion that the case of ML datasets does not give rise to any of these rights. The lack of a human audience voids both the interest of the author in receiving credit and the concern that the author’s dignity would be prejudiced as a result of the modification of the work.<sup>46</sup>

\*\*\*

As discussed, the copyright exposure of ML enterprises is rather high. ML enterprises are vulnerable both to copyright infringement claims of *copyright owners* and to moral rights claims by *authors*. The legal exposure pertains to *each of the numerous works in the dataset*. Furthermore, the remedies for copyright infringement are particularly high, and include, *inter alia*, injunctions and damages without proof of damage in the

---

<sup>42</sup> See Greenman, *supra* note 12, at 823-884; Yehoshua Weissman, *The Moral Right (droit moral) in Copyright Laws*, MECHKAREY MISHPAT (Bar-Ilan Law Studies) G 51 (1989); Gad Tedeschi, *Intellectual Property and Personal Rights*, MISHPATIM J 392 (5740); Pesach, *supra* note 10 (“[] it is necessary to delimit and define the scope of extension of the no-distortion right”); Kim Treiger Bar-Am, *The Moral Right of Integrity, the Defense of Reasonableness, and the Balance Between Them*, 8 ALEI MISHPAT 237, 238 (2010) (“The moral right to the integrity of the work is not an absolute right”). No international harmonization on moral rights exists. See, e.g., § 9(1) of the TRIPS Agreement (Trade-Related Aspects of Intellectual Property Rights) (1994).

<sup>43</sup> §§ 45(a), 50(b) of the Copyright Act: “The author of an artistic work, a dramatic work, a musical work as Personal or a literary work, excepting computer programs, in which copyright Right subsists, shall have moral rights in relation to his work, during the entire period of copyright in that work; however, with regard to a work that is a font, the right as stated in section 46(1) shall not apply.”

<sup>44</sup> See § 46 of the Copyright Act. See also 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, art. 5(3)(a), 2001 O.J. (L 167), 10 (EC) (denies the right of attribution within the framework of the “research exception” when such an indication as said “turns out to be impossible”). (hereinafter: the “Infosoc Directive”). The Directive was applied differently in EU member states. For example, the Italian law obliges to give credit to an author even when copying for research purposes. See (1) Legge diritto d’autore, Art. 70.

<sup>45</sup> See Kimron Case, *supra* note 26; CA 782/87 *Elhanani v. Tel Aviv – Yafo*, SCR 46(3) 529, 537-539 (1992); CA 3422/03 *Krone AG v. Inbar Armored Plastic*, SCR 59(4) 365, 375-376 (2005); CC (Tel Aviv District Court) 1299/04 *Kook v. Sivan In House Ltd.* (Nevo, 13.2.2008); CC (Jerusalem District Court) 8303/06 *Iraki v. Hatib*, para. 5 of the judgment (Nevo, 11.2.2015); Tony Greenman, *The Moral Right – from Droit Moral to Moral Rights*, in *CREATING RIGHTS – READINGS IN THE COPYRIGHT ACT 458* (Michael Birnhack & Guy Pesach Eds., 2008) (“case law does not include many examples of cases in which it was found that there was no obligation to refer to the name of the author under the circumstances of the case”). It is clarified that the said does not constitute an expression of agreement to the broad interpretation of the moral right. A discussion of this issue exceeds the scope of the Opinion.

<sup>46</sup> See also European Parliament, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects – In Depth Analysis*, 2018, [https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL\\_IDA\(2018\)604941\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf), 11 (states regarding the right of integrity that: “The massive amount of materials to be mined would make practically impossible to fulfill this requirement”).

## State of Israel Ministry of Justice

amount of up to NIS 100,000 for *each instance of infringement*.<sup>47</sup> It is therefore crucial to examine whether ML enterprises satisfy the conditions of copyright limitations under copyright law.

### 2. COPYRIGHT LIMITATIONS

Alongside section 11 discussed above, the Statute sets forth limitations to copyright owners' exclusive rights. These limitations, which are enumerated in sections 19-32 of the Statute, permit the use of copyrighted works without rightsowners' permission and without paying consideration for the use.<sup>48</sup> As shown below, three of these exceptions might apply to ML: fair use (section 19), incidental use (section 22) and transient copying (section 26). As we show below, the fair use doctrine is the primary framework for this analysis, yet the other provisions might also apply in relevant cases.

#### A. Fair Use

The fair use doctrine comprises the major balancing mechanism between the interest of protection of works and the public interest in increasing access to the works.<sup>49</sup> The fair use doctrine in Israel is defined in section 19 of the Statute:

**“19. (a) Fair use of a work is permitted for purposes such as: private study, research, criticism, review, journalistic reporting, quotation, or instruction and examination by an educational institution.  
(b) In determining whether a use made of a work is fair within the meaning of this section the factors to be considered shall include, *inter alia*, all of the following:  
(1) The purpose and character of the use;  
(2) The character of the work used;**

---

<sup>47</sup> See § 56 of the Copyright Act.

<sup>48</sup> See §§ 19-32 of the Copyright Act.

<sup>49</sup> See the *Premier League Case*, *supra* note 4, paras. 17-18 in the judgment of Deputy President Rivlin: “This is a tool [...] that is intended to allow, if necessary, to strike a balance between the aspiration to give an incentive to authors to produce new works, and the aspiration to enrich the diversity of the works in the public sphere [...]”; ORIT FISCHMAN -AFORI, DERIVATIVE WORK IN COPYRIGHT LAWS 332 (2005) (“given the fact that the fair use defense is a tool for the purpose of striking a balance between the right holder and the public interest, naturally there is a dispute on the question whether the desired balance is achieved by it and whether it suffices for the purpose of protecting the public interest”); Niva Elkin-Koren, *Users Rights*, in CREATING RIGHTS: READINGS IN THE COPYRIGHT ACT 327 (Michael Birnhack & Guy Pesach Eds. 2008); Michael Birnhack, *Judicial Snapshots and Fair Use Theory* 5, QUEEN MARRY, J. INTELL. PROP. 264, 264 (2015).

## State of Israel Ministry of Justice

- (3) The scope of the use, quantitatively and qualitatively, in relation to the work as a whole;**
- (4) The impact of the use on the value of the work and its potential market.**
- (c) The Minister may make regulations prescribing conditions under which a use shall be deemed a fair use.”**

Per section 19, the fair use analysis comprises two prongs. First, the use must be made for one of the purposes stated in section 19(a) (or “for purposes such as these”).<sup>50</sup> Second, the use must meet the fair use criteria enumerated in section 19(b).<sup>51</sup>

Begin with section 19(a). A strong argument can be made for the proposition that the creation of ML datasets fall under both “self-learning” and “research,” and, at the least, is made for purposes “such as these” purposes.<sup>52</sup> ML comprises “self-learning” because the machine itself learns using datasets. Indeed, ML is equivalent to the human process of inductive self-learning. Both processes are performed by way of learning from examples. The distinction between human learning and ML concerns only the technical-technological process of learning: while the human brain can learn while randomly coming across examples with no a specific dataset, computers can “learn” only from datasets that are organized in a specific form. Because this distinction merely concerns the technical process of learning, an interpretation of the term “self-learning” as including ML aligns with both

---

<sup>50</sup> The chapter on permitted uses replaced a closed list of permitted uses from the previous law – § 2(1) of the Copyright Act 1911, Law of Israel C 2633 (70). This shift was perceived as reflecting a legislative intent to expand the free use doctrine. *See* Greenman, *supra* note 12, at 327 (“the provisions for the purpose of this matter [the permitted uses] are laid down in principle in chapter D of the Law, titled “Permitted uses.” These provisions are the most material innovation in the Copyright Act compared to the previous law, and they caused, following the legislation of this law, a change in the balance between the copyright and the interests of the users of the work.”); Yehoshua Weissman, *Comparative Reading: Characteristics of the Copyright Act 2007*, in *CREATING RIGHTS: READINGS IN THE COPYRIGHT ACT (5769)* 61, 80 (“...the open list of the ‘fair use’ defense, same as the provision denying the copyright defense from the ideas in the works, reduces significantly the rights of the authors, with special emphasis on the public interest in the free use of cultural works, compared to the interests of the authors in the products of their work”). A similar transition from a closed to an open list was also made in Australia. *See, e.g.,* Emily Hudson, *Implementing Fair Use in Copyright Statute: Lessons from Australia*, 25(3) I.P.J. 201 (2013).

<sup>51</sup> De facto, the obstacle set out in § 19(a) is not dominant in case law (and sometimes is not at all considered). The main analysis concentrates on § 19(b). *See, e.g.,* the Premier League Case, *supra* note 4, paras. 19-20 in the judgment of J. Rivlin (‘skipping’ a § 19(a) examination); CA 7996/11 *Safecom Ltd. v. Ofer Raviv*, para. 37 in the judgment of J. Danziger (Nevo, 18.11.2013) (hereinafter: the “Safecom Case”) (same). *See also* Greenman, Copyright, 406-407 (2008) (“...similar to the U.S. courts – it is necessary to examine the question in one continuous stage, that is intended to examine the fairness of use, when the purpose of use is one of the considerations that will be taken into consideration as part of the first test, that concerns the purpose and character of the use”). *But see* CA 3425/17 *Societe des Produits Nestle et Ors. v. Espresso Club Ltd.* (7.8.2019) (hereinafter: the “Espresso Club Case”), para. 21 in the judgment of J. Hendel (“the language of section 19(a) does not allow to forgo the purpose test as an independent threshold test”).

<sup>52</sup> *See* § 19(a) of the Copyright Act.



## State of Israel Ministry of Justice

the present technological reality and with the Israeli case law that has interpreted the term “self-learning” broadly.<sup>53</sup>

The creation of ML datasets probably also falls under the definition of “research,” both because processes of analysis, selection, sorting and characterization are inherent to dataset creation and because the way computers learn is itself subject to research.<sup>54</sup> Notably, the interpretation of both terms—“research” and “self-learning”—has traditionally been broad and has included commercial uses.<sup>55</sup> In any event, the Section provides an “open” list and permits the use of protected materials also for purposes “*such as* these.” Therefore, even had the creation of the dataset failed to meet a literal-strict definition of “self-learning” or “research” (despite their broad interpretation in case law), it is sufficiently close to them for the purpose of satisfying, in our opinion, the condition set out in section 19(a).

After passing the hurdle of section 19(a), it is time to examine the four fair use factors, listed in Section 19(b). These factors should be examined, “*inter alia*,” for the purpose of “determining whether a use made of a work is fair.”<sup>56</sup> While additional considerations can be examined, the four considerations enumerated in the Statute are fundamental. Below we delineate an analysis of the four fair use factors in the context of ML datasets. As analyzed below, this analysis leads to the conclusion that the frequent cases of ML dataset creation falls under fair use.

### (1) *The purpose and character of the use*

The first fair use factor concerns the “purpose and character of the use.”<sup>57</sup> The *purpose* of use in the case in question is the creation of effective ML datasets. The creation

---

<sup>53</sup> See, e.g., CC (Jerusalem District Court) 8303/06 *Mehola Dance Center Ltd. v. Cohen*, DCR 3(08) 7396 (2008), where the teaching of a dance via its performance before and by students constitutes self-learning.

<sup>54</sup> See also Dan Brown, Lauren Byl & Maura R. Grossman, *Are ML corpora “fair dealing” under Canadian law?* 159 (2021) (hereinafter: “Brown, Byl & Grossman, “fair dealing”) (noting that the copyright exceptions in Canadian law interpret the term “research” broadly, citing *CCH Cdn. Ltd. v. LSUC* (2004), 266 F.T.R. 159 (FC), which states that “‘Research’ must be given a large and liberal interpretation in order to ensure that users’ rights are not unduly constrained”).

<sup>55</sup> See, e.g., Greenman, *supra* note 12, at 414 (“the question that should be asked...[is] whether the author created the material, explained, commented or criticized it, or whether he responded in any other manner that can be considered as “research” with respect to the material itself”). See also CC (Tel Aviv Magistrates Court) 24595/97 *Bass v. Keter Publishing Ltd.*, PM 5762(3) 337, para. 61 in the judgment of Judge Shachar (2002) (finding that a use is considered research because it explains the works’ textual content); CC (Jerusalem Magistrates Court) 8397/98 *Bitton v. Sultan*, MCR 00(2) 9883 (2000), reaffirmed in LCA 1108/02 *Biton v. Sultan* SCR 02(2) 1110 (2002) (recognizing the purpose of research in commercial contexts). The EU ‘Research Exception,’ set out in Article 5(30) of the Directive, was interpreted as a normative framework for the issue of TDM, which overlaps extensively with this issue. See *infra* Part D.

<sup>56</sup> See § 19(b) of the Copyright Act.

<sup>57</sup> See § 19(b)(1) of the Copyright Act.

## State of Israel Ministry of Justice

of such dataset is essential for AI systems to be effective, safe, and competitive. The character of use mainly examines whether the use is made “AS-IS,” or whether it is transformative, i.e., a use that changes the work, its meaning or its context. The more transformative the use, the “fairer” it is considered, based on the assumption that transformative uses enrich the world of creative works and benefits the public.<sup>58</sup>

Analyzing the transformative nature of works is highly complex in the ML context. Technically, works are copied to datasets without substantial modifications.<sup>59</sup> Yet, a holistic examination of the creation processes of datasets shows that such use is as transformative as it gets, because of the intensity of the contextual change of the use. Consider, for example, an autonomous driving system. Such a system may ‘watch’ movies in order to teach the system the ‘object fixation’ principle and educate it to anticipate that a pedestrian who appears behind a tree will reappear, rather than to enjoy the aesthetic quality or the content of the film. Likewise, when an NLP system ‘reads’ texts, it does not intend to analyze its thesis, but rather, for example, to identify parts of speech in sentences.<sup>60</sup> Granted, the composition, lighting and other creative decisions in the film creation process, as well as the diversity and choice of words of the sentences in the original works, are digested by the system. But the use that the system makes eventually with these tools is completely different from the original use.<sup>61</sup>

Clearly, some ML systems are more transformative than others. In particular, systems can be designed to produce outputs that would highly resemble their inputs.<sup>62</sup> This can be done deliberately (for example, when a system is designed to imitate a specific author or a genre),<sup>63</sup> or unintentionally, such as when the dataset is not sufficiently diverse.<sup>64</sup> In such circumstances, the use may not be considered transformative.

---

<sup>58</sup> See, e.g., the Premier League Case, *supra* note 4, in the judgment of Vice President Rivlin (“The approach is that it is easier to recognize transformative use as “fair” because it attains the purpose of the permission – to encourage the work and enrich the cumulative knowledge in society”); Espresso Club Case, *supra* note 51, para. 28 in the judgment of J. Hendel (“the consideration regarding the transformative nature of the work is perceived at present as the most dominant element in case law in the United States, and also in Israeli case law it was recognized as a major consideration when examining the character of use”); Safecom Case, *supra* note 51, para. 37 in the judgment of J. Danziger; LCA 7774/09 *Weinberg v. Weisshof* (Nevo, 28.8.2012) para. 22 (hereinafter: the “Weinberg Case”). See also *Suntrust Bank. v. Houghton Mifflin Co.*, 268 F.3d 1257 (2001) (analyzing the use of *Gone with the Wind* in Alice Randall’s novel ‘The Wind Done Gone’ as transformative in relation to the original work).

<sup>59</sup> See *supra* chapter B.

<sup>60</sup> See also Mark A. Lemley & Bryan Casey, *Fair Learning*, 99(4) TEX. L. REV. 743 (2021).

<sup>61</sup> See also *Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1176 (9th Cir. 2007) (finding that the inclusion of full thumbnails in search engines is considered transformative due to the different functionality of the use).

<sup>62</sup> Thus, for example, in a class action that was recently filed in the United States in the case of GitHub Copilot – Microsoft’s AI tool for writing code, it was argued, *inter alia*, that the machine output can produce software segments that are identical to the software segments that were included in the dataset. See Complaint, Class action, *J. DOE et al. v. Github, Inc. et al.*, Case N.D. Cal. (filed 11.3.22). Available at: [https://githubcopilotlitigation.com/pdf/1-0-github\\_complaint.pdf](https://githubcopilotlitigation.com/pdf/1-0-github_complaint.pdf).

<sup>63</sup> See, e.g., *id.* see also *supra* notes 136-137.

<sup>64</sup> The use of insufficiently diverse datasets can stem from different reasons, such as the existence of partial data and limited access to existing data. Nondiverse datasets creates different problems besides copyright

## State of Israel Ministry of Justice

Finally, the first fair use factor also considers whether the use is commercial. A commercial nature of the use weighs against fair use,<sup>65</sup> although it does not automatically deny the application of the doctrine, especially when the use is transformative.<sup>66</sup>

In light of the above analysis, we are of the opinion that the paradigmatic case of ML dataset meets the criterion laid down in section 19(b)(1), given its societal value and transformative nature. Nevertheless, in situations where the enterprise is not transformative, and in particular when it is also commercial, section 19(b)(1) may weigh against finding of fair use.

### (2) *The character of the work used*

The second fair use consideration concerns the nature of the work used: the type of the work, and to what degree this work is at the core of copyright law. Thus, for example, the use of a literary or musical work might be examined more rigorously compared to the use of a broadcast or a recording. In addition, the use of fact-based or research works will be classified more easily as fair compared to works that are the figment of the author's imagination. Largely, this consideration is considered secondary among the fair use criteria.<sup>67</sup> In any event, there is an inherent difficulty in evaluating it categorically, considering the endless applications of ML and AI ventures.

---

laws, from inferior effectiveness of the system and its versatility to biases against groups that are not represented in the dataset. *See, e.g.*, Levendowski, *supra* note 36; Kirsten Lloyd, *Bias Amplification in Artificial Intelligence Systems*, 2 (2018). Available at <https://doi.org/10.48550/arXiv.1809.07842>; Bert Heinrichs, *Discrimination in the Age of Artificial Intelligence*, 37 *AI & SOC'Y* 143, 150-151 (2022).

<sup>65</sup> *See, e.g.*, LCA 92/2687 *Geva v. Walt Disney Company*, SCR 48(1) 251, 271 (1993), p. 276; the Premier League Case, *supra* note 4, para. 20 in the judgment of J. Rivlin.

<sup>66</sup> *See, e.g.*, the Premier League Case, *id.* (“the mere existence of commercial use does not deny the approximately of the fair use defense. However, it is customary to say that to the extent that this is indeed commercial use, the argument regarding fair use weakens”); Espresso Club Case, *supra* note 51, paras. 30-31 (“denying the defense from commercial use, certainly for the purpose of making profit, might void extensive parts from the fair use arrangement, and affect the important social purposes for which it is intended”). *See also* *Campbell v. Acuff-Rose Music Inc.*, 114 S.Ct. 1164, 1176 (1994); *Authors Guild v. Google, Inc.*, 804 F.3d 202, 219 (2d Cir. 2015) (“[there is] no reason . . . why [a defendant’s] overall profit motivation should prevail as a reason for denying fair use over its highly convincing transformative purpose, together with the absence of significant substitutive competition, as reasons for granting fair use”) (hereinafter: the “Google Books Case”).

<sup>67</sup> *See, e.g.*, the Premier League Case, in which this factor was not even examined. *See also* the Espresso Club Case, *supra* note 51, para. 34 in the judgment of J. Hendel (“The consideration of the character of the work will occupy a dominant place only in unique circumstances”). Empirical analyses of court cases in the United States show a similar trend. *See, e.g.*, Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 *U. PA. L. REV.* 549 (2008); Neil W. Netanel, *Making Sense of Fair Use*, 15 *LEWIS & CLARK L. REV.* 715 (2011); Mathew Sag, *Predicting Fair Use*, 73 *OHIO ST. L. J.* 47 (2012).

## State of Israel Ministry of Justice

### *(3) The scope of the use quantitatively and qualitatively*

The third fair use factor concerns the scope of the use qualitatively and quantitatively.<sup>68</sup> The idea is to examine what and how much was taken from the original work for the purpose of the use. In the ML context, works are typically reproduced and included in the datasets in full—supposedly an indication against finding of fair use. Nevertheless, sometimes (albeit not always) works are copied to datasets not for the purpose of using their creative aspects, but to enable access to *unprotected* elements that are included in the work, such as facts, ideas, syntactic structures, and the like.<sup>69</sup> The work is copied in full only because a computer, unlike humans, cannot access unprotected materials without first copying *the entire work* into a readable dataset.<sup>70</sup> This is a technical limitation that stems from the inherent architecture of learning systems, and an interpretation that aims to enable ML must take it into consideration. Consequently, the *scope of the use* factor should not turn on the reproduction of the work in full, but also consider the elements that were actually used in the learning process. When learning is performed mainly from unprotected elements, the third factor, in our opinion, should add further support to the finding of fair use.

### *(4) The impact of the use on the value of the work and its potential market*

The fourth fair use consideration analyzes the impact of the use on the actual and potential markets of the work. As explored below, in our view this analysis tips the scale in favor of finding of fair use in the ML context.

Consider first the actual market of the works. The expansion of machines' access to works will not harm existing markets of copyright owners, because *such markets are nowhere to be found*. Indeed, even if ML enterprises intended to purchase licenses for each of the works in the dataset, doing so would be practically impossible. No platform offers such licenses, while the scope of the required works, their geographic distribution, and the lack of registration of rights in the works eliminate any possibility to obtain licenses directly from rightsholders.<sup>71</sup> Notably, in this context too, the more diverse the dataset, the stronger the conclusion regarding the lack of a market of licenses. A dataset that is composed of works of a single author may very well be able to negotiate a license.

---

<sup>68</sup> See *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 556 (1985) (ruling that copying 300 words from a biography of 400 pages was considered substantial in terms of quality of the work).

<sup>69</sup> See § 5 of the Copyright Act. See also Daniel Gervais (2019) *The Machine as Author* 24 REV. L. IOWA 105 (“TDM is looking, if anything, for ideas embedded in copyright works.”).

<sup>70</sup> See also *supra* notes 16-17 and accompanying text.

<sup>71</sup> See also the Google Books Case, *supra* note 66, in which the court ruled that the Google books project did not harm rightsholders and in fact contributed to them by the publication of their works.

## State of Israel Ministry of Justice

In addition to an analysis of market impact in the static sense, namely, determining whether such a market exists—the fourth factor requires an analysis of the impact on the market dynamically, namely, the impact on potential markets for works.<sup>72</sup> After all, finding fair use in the ML context based on the dearth of markets at present may impede the creation of a market for this use in the future. Nevertheless, we are of the opinion that dynamic analysis also leads to the conclusion that the inclusion of works in ML datasets constitutes fair use. The reason for this is that inherent market failures give rise to a very low probability that an effective licensing market will be created for such uses. Such market failures include information problems and coordination problems that stem, *inter alia*, from the absence of registration of rights in works, and massive transaction costs that are the product, *inter alia*, of the distribution of works (and the laws governing them) over the globe. To date such market failures, among others, prevented the development of efficient markets in numerous contexts relating to copyrights in the digital world.<sup>73</sup> These market failures are prone to thwart an efficient licensing market for ML purposes.

What is more, finding that the fair use doctrine applies to ML does not necessarily impede the creation of such a market.<sup>74</sup> On the contrary, the fair use doctrine might actually incentivize the creation of a market. To the extent that an efficient market for licensing of works for ML is created, the fourth factor might reverse its direction and weaken the fair use argument in the certain contexts in which such markets are created. For now, however, the fair use doctrine allows the AI industry to thrive until such markets emerge, noting the inherent difficulties in the creation of such markets.<sup>75</sup>

Furthermore, finding no fair use in this context can result in negative competitive effects. First, high transaction costs can create severe barriers to entry to the AI sphere for early-stage companies. Indeed, resourceful incumbents can leap over the legal uncertainty hurdle more easily, for two reasons. First, incumbents have accumulated vast data over the years. Google, for example, was able to compose a vast dataset of email correspondences of its millions of users over decades for the training of its NLP system. Early-stage ventures possess no such data, and will need to obtain data elsewhere or use synthetic data, which might affect the quality of their products. Second, facing legal uncertainty, resourceful companies can obtain a license to use works even when such a license is not clearly required. Obtaining licenses for uses that can be performed without a license by virtue of

---

<sup>72</sup> In the beginning of the century courts in the United States considered the idea that the mere potential of a consumer market should deny fair use. See *Princeton Univ. Press v. Mich. Document Servs., Inc.*, 99 F.3d 1381, 1388 (6th Cir. 1996); *Am. Geophysical Union v. Texaco, Inc.*, 60 F.3d 913, 930–31 (2d Cir. 1994) (“[A]n unauthorized use should be considered ‘less fair’ when there is a ready market or means to pay for the use.”) This position spurred intense criticism in the literature. See James Gibson, *Risk Aversion and Rights Accretion in Intellectual Property Law*, 116 YALE L.J. 882, 931–35 (2007); Wendy J. Gordon, *The ‘Why’ of Markets: Fair Use and Circularity*, POCKET PART 371 (2007), <http://yalelawjournal.org/2007/4/25/gordon.html>; Niva Elkin-Koren, *Self-Regulation of Copyright in the Age of Information*, 319 ALEI MISHPAT(B) 341 (2002).

<sup>73</sup> See, e.g., Lital Helman, *Fair Trade Copyright*, 36(2) COLUM. J. L & ARTS 157.

<sup>74</sup> See, e.g., *supra* note 76 and accompanying text.

<sup>75</sup> There is inherent circularity in the fourth fair use factor. The absence of markets for the works leads to finding of fair use, but finding of fair use may decrease the incentive to create a market. It may thus be hard to escape a normative decision on whether such a market is *desired*. It is hereby clarified that this Opinion deliberately avoids making such a normative decision.

## State of Israel Ministry of Justice

fair use is customary for risk-averse resourceful players.<sup>76</sup> Obviously, startup companies cannot afford excessive licensing. Moreover, defensive licensing practices of incumbents might create negative externalities on startups, because such practices may weaken startups' reliance on fair use. Startups' fair use claim will be weakened *de facto*, because rightsholders will *expect* to receive payment for such uses, as well as *de jure*, because the existence of a market would affect the decision whether the use of a work is considered fair.<sup>77</sup> As a result of a weak fair use claim, AI may 'migrate' to incumbents' domains, generating costs for innovation and entrepreneurship, around the world generally and in Israel in particular.<sup>78</sup>

A second competitive concern concerns copyright ownership. Determining that ML dataset composition requires a license can increase the centralization in content markets. Such a conclusion could strengthened the market status of centralized content entities at the expense of individual authors.<sup>79</sup> Indeed, deciding that dataset creation requires a license would generate a powerful incentive to collect materials for ML datasets from massive content aggregators, such as large publishers, radio stations, collective management organizations of copyrights and other rightsholders that are easy to locate and transact with. Such an incentive could intensify centralization and may also adversely affect the diversification of datasets.

Compared to these prices, the damage to rightsowners from adding their works to ML datasets is infinitesimal, if at all. First, as discussed, no efficient licensing market exists for such uses, and therefore, even if we were to conclude that such uses are not considered fair, *ML enterprises would not have been able to purchase a license for their operations*, except perhaps from major content owners or rights management organizations, in certain circumstances. Even had an efficient licensing market for ML existed, the price for licensing each work's would have reflected its marginal value in the dataset. When striking a balance between the low, if any, profit a license could have produced for rightsowners as a result of a narrow interpretation to the fair use doctrine and the substantial economic and competitive gain that a broad application of the doctrine could generate, the latter has the upper hand.

\*\*\*

To sum up, in most cases, the fair use factors cumulatively lead to the conclusion that the doctrine covers the use of copyrighted materials in the ML context. Thus, *the purpose and character of the use* is typically transformative and done for a worthy cause,

---

<sup>76</sup> See Gibson, *supra* note 74, at 931–35; Wendy J. Gordon, *The 'Why' of Markets: Fair Use and Circularity*, POCKET PART 371 (2007), <http://yalelawjournal.org/2007/4/25/gordon.html>; Wendy J. Gordon & Daniel Bahls, *The Public's Right to Fair Use: Amending Section 107 to Avoid the "Fared Use" Fallacy*, 2007 UTAH L. REV. 619 (2007); Elkin-Koren, *supra* note 73.

<sup>77</sup> See §19(b)(4) of the Copyright Act.

<sup>78</sup> See JOSEPH SCHUMPETER, *BUSINESS CYCLES* 93 (1939) (noting that early-stage companies are more innovative than established ones).

<sup>79</sup> Centralization in content markets is a well-known phenomenon. See, e.g., Lior Baruch, Maayan Perl & Niva Elkin-Koren, *Competition and Power Gaps in the Digital Music Market: Rerouting*, 331 REGULATION STUDIES D (2021).

## State of Israel Ministry of Justice

albeit sometimes commercial; the *character of the work* differs from one case to another, and cannot be categorically addressed; the *scope of use* points in the direction of fair use in most cases, in particular when despite the reproduction of the work in full, the learning is done from its noncopyrighted parts; and *the impact on the market of the work* is negligible at best, both based on the present situation and in light of a structural analysis of the content markets in the online arena.

In addition to the four fair use criteria, Israeli case law typically examines as part of the question of the fair use whether attribution to the author was made (this Opinion should not be read to endorse this added question).<sup>80</sup> Based on the analysis provided in this Opinion, there is no point in giving attribution to authors in front of a machine.<sup>81</sup> Therefore, the fact that credit is not given to the author in the process of creation of the dataset does not alter our conclusion.<sup>82</sup>

At the same time, the context of ML datasets calls for exploring a new factor: the *diversity* of the dataset.<sup>83</sup> A dataset that comprises diverse content is more likely to be fair than one that is composed by works of a single author from which the system learns the specific style of the author. We posit that the more diverse the dataset, the more likely the finding of fair use. Most datasets aspire for diversity, in order to maximize the potential advantages and use-cases of the system. When datasets are not diversified, this consideration would weigh against the finding of fair use.

The conclusion that ML datasets typically constitutes fair use aligns well also with the literature regarding fair use. In fact, the issue in question well exemplifies some of the main theoretical justifications for the fair use doctrine. A major justification for the fair use doctrine stems from the concept of productivity. Under this approach, fair use facilitates desirable uses that could not occur without it.<sup>84</sup> Indeed, market failures, mostly high transaction costs and ‘holdup problems’ would have otherwise frustrated desirable uses in the works.<sup>85</sup> Another justification for fair use concerns striking a balance between copyrights and users’ interests, which was discussed in detail above.<sup>86</sup>

As to productivity, the ability to use creative works will boost the incentives to engage in AI, especially for small and new firms. Absent the fair use doctrine, AI enterprises will face prohibitive market failures. First, obtaining licenses from rightsowners in millions of works, who may reside in different places around the globe produces high

---

<sup>80</sup> See, e.g., CA 8117/03 *Inbar v. Yaakov*, para. 23 in the judgment of J. Naor (Nevo, 16.1.2006); CA (Tel Aviv District Court) 3038/02 *Zoom Communication (1992) Ltd. v. Israeli Educational Television*, para. 5(f) in the judgment (Nevo, 29.4.2007), CC (Jerusalem Magistrates Court) 7036-09 *Rachmani v. Israeli News Company Ltd.* (Nevo, 9.10.2011). See also Birnhack, *supra* note 11; Lital Helman, *Session IV: Fair Use and Other Exceptions*, 40 COLUM. J.L. & Arts 395. But see CC (Jerusalem Magistrates Court) 8211-09 *Forges v. Western Galilee High School* (Nevo, 27.7.2011).

<sup>81</sup> See, *supra* notes 45-46, and accompanying text.

<sup>82</sup> It will be interesting to follow the United States case of GitHub Copilot. See *supra* note 62.

<sup>83</sup> See § 19(b) of the Copyright Act (providing that the list of considerations is an open list).

<sup>84</sup> See Dafna Levinson-Zamir, *The ‘Fair Use’ Defense in Copyright*, 16 MISHPATIM 430, 432 (1985).

<sup>85</sup> See generally Gordon & Bahls, *supra* note 78 (theorizing that fair use aims to tackle market failures).

<sup>86</sup> *Id.* See also ORIT FISCHMAN-AFORI, *DERIVATIVE WORK IN COPYRIGHT LAWS* 331 (2005).

## State of Israel Ministry of Justice

transaction costs that drastically reduce the feasibility of AI enterprises.<sup>87</sup> The creation of an effective dataset would require finding and negotiating with each copyright owner, while to satisfy the conditions regarding moral rights they will have to locate the author herself (though we posit that moral rights are not violated in this context).<sup>88</sup> Furthermore, works may comprise various copyrights. For example, to include music files in a dataset, negotiations will have to be pursued from the copyright owners in the lyrics, in the melodies, and in the records, and perhaps even from performers and broadcasters.<sup>89</sup> The costs involved in finding the many copyright owners will be fully imposed on the ML ventures. Neither copyright owners whose works are made available to the public nor the platforms that display the works have any obligation to specify the status of rights in the works. Indeed, works online rarely specify the works' copyright status and the Law grants no protection to users who rely on incorrect information of the author's identity.<sup>90</sup>

The concern about prohibitive transaction costs is related to another market failure, known as 'holdup problem:' fragmentation of exclusive rights among a vast number of rightsholders allows each rightsowner to frustrate the project, or, at least, to severely delay it.<sup>91</sup> Granted, the power of a single rightsowner to impede the entire project is limited because each individual work in the dataset is substitutable. Yet, competitive constraints in the world of entrepreneurship, together with the need to meet ambitious milestones in order to raise funds and meet targets, mean that delays—which every single rightsholder *can* impose—may very well thwart the entire project.<sup>92</sup> Most importantly, the fact that these concerns are market failures means that *failure to apply copyright limitations will not result in payment to copyright owners but rather inhibit AI enterprises altogether.*

\*\*\*

Before we conclude this issue, it is necessary to add a few more words. Conceptually, the fair use defense examines unauthorized uses ad-hoc rather than categorically. Fair use decisions are typically made retroactively, after unauthorized use of copyrighted content has been made and the user raises the fair use argument in a litigation (or pre-litigation) process ex-post. The argument is then examined ad-hoc, based on the concrete use made. An ex-ante statement that an entire category of uses falls under the fair use doctrine might appear somewhat anomalous.

---

<sup>87</sup> Wendy J. Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors*, 82 COLUM. L. REV. 1600 (1982).

<sup>88</sup> See *supra* note 41 and accompanying text.

<sup>89</sup> A discussion of performers' and broadcasters' rights exceeds the scope of this Opinion. See generally § 4 of the Copyright Act; Broadcasters' and Performers' Rights Law 5744-1984.

<sup>90</sup> The 'innocent infringer' defense, set out in § 58 of the Copyright Act, protects only users who did not know or who could not have known that copyright persists in the work.

<sup>91</sup> See Ben Depoorter & Francesco Parisi, *Fair Use and Copyright Protection: A Price Theory Explanation*, 21 INT'L REV. L. & ECON. 453 (2002). For example, following a case that ruled that each of the thousands past athletes have the right to prevent the creation of a football video game (*Keller v. Elec. Arts, Inc.*, 724 F.3d 1268, 1269 (9th Cir. 2013)) the game producer abandoned the project altogether.

<sup>92</sup> See Lital Helman, *Innovation Policy and the Valley of Death*, SMU L. REV. (forthcoming 2023), available at <https://ssrn.com/abstract=4262740> (discussing the vitality of timing in entrepreneurial projects).



## State of Israel Ministry of Justice

Yet, the analysis proposed here, namely that the creation of ML datasets constitutes in most cases fair use, does not exceed the bounds of the doctrine. On the contrary, when there is a dominant common denominator for a defined category of uses, the ex-ante interpretation of the fair use doctrine can enhance certainty for market players on both sides, reduce litigation costs and promote efficiency.<sup>93</sup>

Moreover, the analysis provided in this Opinion does not take from the fair use doctrine its concrete ad hoc essence. There can well be circumstances in which an ad-hoc analysis of a specific use of copyrighted works for ML will *not* be considered fair use. Consider, for example, a dataset compiled exclusively of a specific photographer's works that is used to train the machine to mimic that photographer's style. The venture then sells the works of its generative AI for a lower price, or competes with this photographer on a large photography project. Similarly, imagine a machine that produces summaries of textbooks after being trained on the full texts of the books and harms the market of the original textbooks. Such cases may very well be outside of the safe harbor portrayed in this Opinion.

Not only might specific cases exceed the scope of the fair use doctrine, but the underlying AI technologies and markets might undergo extreme transformation. For example, novel ML technologies might skip the need to copy the work in the first place or evolve in other directions. Systems that will allow micropayments or other monetization of datasets production may emerge. Indeed, the ML realm evolves rapidly and the issues that it raises are dynamic and diverse. This Opinion acknowledges the inherent dynamic nature of this sphere and leaves intentionally flexible margins in the application of the fair use doctrine to ML.

### *B. Incidental Use of a Work*

Another copyright limitation that is applicable to the case in question concerns incidental use of works. Section 22 of the Statute permits including works in another work of photography, a cinematographic work or a sound recording, as follows:

**“An incidental use of a work by way of including it in a photographic work, in an audiovisual work or in a sound recording, as well as the use of a such work in**

---

<sup>93</sup> An example of the application of fair use to a category of uses concerns the issue of private uses. Israeli law does not grant specific defense for private use. There are also circumstances in which private use may not be considered fair. (*See, e.g.,* the Performers' and Creators' Rights (Recordings) (Legislative Amendments) Act 1996, which defines a compensation mechanism for private uses in certain context (the arrangement is known as the 'Empty Tapes Act')). Yet, private uses are typically considered fair use, in the United States as well. *See, e.g., Sony Corp. of America v. Universal City Studios, Inc*, 464 U.S. 417 (1984) (holding private noncommercial use to be fair); Frances Grodzinsky & Maria C. Bottis, *Private Use as Fair Use: Is It Fair?* 37(2) COMP. & SOC. 11, 12 (2007) (“Private use in US law is a sub-set of fair use.”).

## State of Israel Ministry of Justice

**which the work was thus incidentally contained, is permitted; for the purpose of this matter, the deliberate inclusion of a musical work, including its accompanying lyrics, or of a sound recording embodying such musical work, in another work, shall not be deemed to be an incidental use.”**

When the product of the machine is the creation of a photograph, an audiovisual work or in a sound recording, it may be possible to regard the reproduction of works into the ML dataset as incidental use. This is because each individual work in the dataset is a negligible component in learning, and does not constitute a major addition to the dataset or the system functionality.<sup>94</sup>

In fact, the ‘incidental’ nature of the use is enhanced in the ML context. The assumption in section 22 is that the protected work is included—albeit not in a material way—in the new work.<sup>95</sup> In the ML context, the work is included in the dataset, but the dataset is only a raw material for the system. The final product of the machine does not include the work at all.

This conclusion also aligns with the U.S. case law. There, in the *Baker* Case, the court held that use of copyrighted works that is incidental and necessary for non-infringing purposes is permitted.<sup>96</sup> The *Baker* case concerned reverse engineering of software, and it may well be applicable to the ML context, in cases where the system’s products are non-infringing.<sup>97</sup>

Two qualifications to the application of the incidental use doctrine in the ML context apply. First, the Statute limits the incidental use doctrine to specific types of works, and excludes cases where the use deliberately includes musical works or sound recordings. These exceptions would apply to ML, meaning that the doctrine will only apply to generative AI that produces works that are photographic, audiovisual or sound recordings. Yet the ‘deliberation’ exception should not be construed too broadly in the ML context. Technically, all the works in the dataset were deliberately included. However, posit that as long as the dataset is diverse, and as long as the system did not deliberately selected works of single authors, the system should be eligible to enjoy the doctrine.

The second qualification of section 22 eligibility concerns cases in which the use is in fact not incidental. A good example includes generative AI that deliberately uses works of specific authors to mimic their style. Nothing in the selection and use of these works can

---

<sup>94</sup> See Greenman, *supra* note 12, at 340-343. See also CC 53689-10-17 *Bardugo v. D. Eitan/R. Lahav Rig Architects and City Planners Ltd. et al.*, para. 55 (16.8.20).

<sup>95</sup> See Explanatory Notes to the Copyright Act: “...in principle, there could be copyright infringement in music that is played in the background when filming a public event, or when filming a public event or a photo that is hung on the wall in the place where a film or a news report is filmed. Even though these actions can be considered “*de-minimis*”, it is proposed to clarify that they are not infringing (Bill 6768 no. 196, p. 1116).

<sup>96</sup> See *Baker v. Selden*, 101 U.S. 99, 104 (1880).

<sup>97</sup> See Pamela Samuelson, *The Story of Baker v. Selden: Sharpening the Distinction between Authorship and Invention*, INTELLECTUAL PROPERTY STORIES 159-193 (2005).

## State of Israel Ministry of Justice

be considered incidental. Another example may concern uses of the dataset itself or the works that it contains. For example, a business model revolves around creating and selling datasets cannot be deemed as making an incidental use in the works that compile the datasets.

### *C. Transient Copying*

Finally, the third doctrine that can permit the use of copyrighted works for ML applies in cases where the copying of the works to the dataset is temporary. The permission to perform temporary copying is defined in section 26 of the Statute as follows:

**“26. The transient copying, including incidental copying, of a work, is permitted if such is an integral part of a technological process whose only purpose is to enable transmission of a work as between two parties, through a communications network, by an intermediary entity, or to enable any other lawful use of the work, provided the said copy does not have significant economic value in itself.”**

The creation of ML datasets can fall under the final option in the Statute, “to enable any other lawful use of the work.”<sup>98</sup> This condition will be satisfied when besides the copying of the materials to the dataset, the use of the work is “lawful,” such as when the access to the works was lawful. In addition, the work must not have a significant economic value.<sup>99</sup> In the typical case of ML datasets, this condition is technically met because each individual work in the dataset has infinitesimal value. Nevertheless, *the dataset itself* could have a significant economic value, especially after it was “cleaned,” “labeled” and customized for machine reading. The independent economic value of the dataset will, in our view disqualify it from section 26 protection. Anyway, for a dataset to be eligible to the section 26 protection, it must be deleted after the use.

Admittedly, applying section 26 to ML datasets entails a broad interpretation of this provision, which was enacted with software that makes temporary copies during ordinary operations in mind. Yet, there was of course no legislative intent to *exclude* ML from protection, as such a technology was not yet invented.<sup>100</sup> Notably, the EU finds the transient

---

<sup>98</sup> See § 26 of the Copyright Act.

<sup>99</sup> The accepted interpretation of the term economic value requires not only a *potential* economic value for the work in theory but a *realization* of this value as well. See Greenman, *supra* note 12, at 365.

<sup>100</sup> Technological innovation often requires a reexamination of copyright law. See, e.g., MCA (Tel Aviv District Court) 11646/08 *The Football Association Premier League Ltd. v. John Doe*, para. 1 (Nevo, 2.9.2009, reversed in the Supreme Court, in CA 9183/09, *supra* note 4) (“The digital revolution and the internet changed our lives. [...] This change requires new balances that will take into consideration the rights of the

## State of Israel Ministry of Justice

use exception, set forth in Article 5(1) of the Directive to be an adequate framework for this issue.<sup>101</sup> The transient copying exception under the EU law applies in circumstances in which the copying is transient or incidental to a technological process that is pursued for non-infringing purposes and that does not have an independent economic value.<sup>102</sup>

In conclusion, while temporary datasets are rare at present, due to the expenses involved in the process for the dataset creation, in circumstances in which the dataset is deleted after use section 26 can protect dataset creators in copyright infringement cases. The section will be unapplicable for enterprises that keep the dataset or for companies that create datasets for AI enterprises. These market players exploit the economic value of the dataset itself and do not make transient use in this dataset.

\*\*\*

The interpretation offered in this Parts aims to strike a balance between ML and copyright in the context of ML datasets that contain copyrighted works. Under the analysis proposed in this Part, the fair use doctrine shall apply to the use of copyrighted works for creating ML dataset. Likewise, the incidental use and transient copying doctrines will also protect this activity in certain circumstances, as discussed above. Yet, certain situations justify a shift of the balance in the direction of enhanced copyright protection. Some such situations are discussed in this Opinion, while others are unforeseeable in the present time. This interpretation will allow the AI market to thrive while causing minimum harm – if any – to the copyright owners. The next Part will portray the law in other countries on this issue.

### D. COMPARATIVE LAW

All legal systems around the world endeavor to strike a balance between copyrights and the interests that inhere in copyright limitations.<sup>103</sup> Legal systems are also well aware of the unique challenges that arise from the tension between copyright and AI in general,

---

public in the consumption of culture and participation in the cultural discourse...”; *Telran Case*, *supra* note 9 (regarding the question whether a device that ‘bypasses’ Technological Protection Measures is infringing).

<sup>101</sup> See Article 5(a) of the Infosoc Directive; *see also* European Parliament, *supra* note 46, at 11.

<sup>102</sup> See CJEU, C-360-13, *Public Relations Consultants Association* (5 June 2014), ECLI:EU:C:2014:1195, §§ 43, 50 (“[an act of reproduction is incidental] if it neither exists independently of, nor has a purpose independent of, the technological process of which it forms part”). *See also* *infra* Part D.

<sup>103</sup> Despite the general support of a balanced approach, legal systems differ on the desirable balancing mechanisms. A common balancing mechanism is the fair dealing doctrine, which originated in the UK and is now prevalent around the world. This mechanism strikes a balance by setting a closed list of permitted uses in copyrighted works (sometimes construed broadly). Another dominant approach is the fair use approach that originated in the United States and has spread to Israel, Singapore, Poland, the Philippines, and South Korea. The fair use doctrine expands the range of permitted uses in copyrighted works and includes criteria for, in lieu of a closed list of, permitted uses. Other legal systems lay down other balancing mechanisms, such as specific statutory exceptions with neither a fair use nor fair dealing mechanism. Legal systems that include fair use or fair dealing can include in addition defined statutory exceptions.

## State of Israel Ministry of Justice

and the question of datasets used for ML in particular.<sup>104</sup> As analyzed below, policies around the world is consistent with this Opinion and allow the use of copyrighted works for ML datasets, in varying levels of flexibility.

Legal systems applied two main approaches for the regulation of ML datasets. Some legal systems opted to regulate the issue by creating specific provisions for the ML context. Such regulation has typically been formalized within the framework of the Text & Data Mining (TDM) exception. The TDM exception preceded ML. The exception was designed to regulate the automated collection and analysis in the digital space. To date, the UK, Japan, Singapore, and the EU have implemented this approach, and Canada and South Korea are considering to follow suit.<sup>105</sup> Conversely, the United States, currently the global leader in ML technologies and markets, has avoided statutory exceptions, and instead has relied on the fair use doctrine for the purpose of striking the desirable balance. That is also the approach taken in this Opinion. In the following discussion, we briefly describe the extant law on the matter in the legal systems that considered this issue, and the advantages and drawbacks of each strategy. Concisely, the key advantage of the specific regulation strategy is enhanced certainty—at least for some time, while the main advantage of the second strategy lies in its flexibility and adaptability to technological, legal and economic developments.

*Japan* pioneered the creation of a specific ML copyright exception. Like most copyright regimes, the Japanese copyright law includes no fair use provision. Instead, the Japanese Copyright Act delineates a closed list of permitted uses.<sup>106</sup> This list includes, *inter alia*, uses of works in political speeches, displays of artworks by their owners, adaptations of a works for disabled people and various other uses.<sup>107</sup> Japan became the first country to introduce TDM, and in 2018 it has expanded the exception to allow ML to operate without obtaining a license, without paying to rightsowners and without giving credit to the author.<sup>108</sup> The only exception to this permission concerns situations in which the use of the work causes unreasonable harm to the economic interest of the author in light of the nature or the purpose of the work or the circumstances in which the work was used.<sup>109</sup>

---

<sup>104</sup> See, e.g., OECD Council, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (Adopted on May 22, 2019; 2021), available at <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>.

<sup>105</sup> See, e.g., Seung Hoon Park, *Copyright Issues for AI and Deep Learning Services: A Comparison of U.S., South Korean, and Japanese Law*, JTIP BLOG (May 28, 2021), <https://jtip.law.northwestern.edu/2021/05/28/copyright-issues-for-ai-and-deep-learning-services-a-comparison-of-u-s-south-korean-and-japanese-law/> (noting that South Korea is considers a ML exception).

<sup>106</sup> See Chosakuken Ha [Japanese Copyright Act], Law No. 48 of 1970, art. 30-4, translated in Chosakuken Kankei Horei Deta Besu [Copyright-Related Law Database] (Copyright Research & Information Center (CRIC)), available at <https://www.cric.or.jp/english/clj/cl2.html> (hereinafter: “Japanese Copyright Act”).

<sup>107</sup> *Id.* These uses strongly resemble the list of permitted uses set out in the Israeli Act alongside fair use. *cf.* §§ 18-32 of the Copyright Act.

<sup>108</sup> See § 4-30 of the Japanese Copyright Act, *supra* note 106.

<sup>109</sup> *Id.*

## State of Israel Ministry of Justice

A year later, in 2019, *the EU* adopted the DSM directive and laid down two new TDM provisions.<sup>110</sup> The first provision, set forth in Article 3 of the Directive, exempts from copyright liability reproduction and extraction made by research organizations and cultural heritage institutions that carry out TDM for the purposes of scientific research.<sup>111</sup> The second provision, set forth in Article 4 of the Directive, mirrors Article 3 with two major differences: it encompasses a much broader class of users, but it is considerably narrower in scope. Thus, under Article 4, copyright exemption is extended to all types of users who perform TDM for any type of use. But users can only retain the copies “for as long as is necessary for the purposes of text and data mining,” namely copyrighted works must be deleted after the TDM operation. Article 4 also allows rightsholders to exclude their works from the arrangement (opt out).<sup>112</sup> The TDM exceptions have not implemented in all EU member states thus far, and were sometimes adopted narrowly.<sup>113</sup> As noted above, in addition to the TDM exceptions, the EU laid out an exception regarding transient or incidental uses of works, which might also be useful in this context.<sup>114</sup>

The third country that adopted designated arrangements for ML was *Singapore*, at the end of 2021<sup>115</sup>. The adoption of this approach by Singapore is interesting because unlike the other members in the TDM exceptions “club,” the Singaporean copyright law actually includes a fair use provision. The Singaporean fair use section is very similar to its Israeli and American counterpart, and could have been used for the purpose of striking a balance on this issue. Nevertheless, Singapore preferred to enact a specific statutory exception for the creation of ML datasets, in order to increase certainty for the AI industry and promote

---

<sup>110</sup> See 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, Art. 3-4, 2019 O.J. (L 130) 92 (EU) (hereinafter the “DSM Directive”). See also Joao Pedro Quintais, *Rethinking Normal Exploitation: Enabling Online Limitations in EU Copyright Law*, AMI No. 6, at 197-205 (2017).

<sup>111</sup> See Article 3 of the DSM Directive.

<sup>112</sup> See Article 4 of the DSM Directive.

<sup>113</sup> In a number of countries, such as Finland, the arrangement was not adopted as of the date of this Opinion. See Sofia Wang, *Implementation of the DSM Directive is progressing in Finland — what changes will the directive bring to the Finnish copyright legislation?*, BIRD & BIRD (June 20, 2022), available at <https://www.twobirds.com/en/insights/2022/finland/implementation-of-the-dsm-directive-is-progressing-in-finland-what-changes-will-the-directive-bring>. For a narrower version of the exception, see, e.g., the French law - Art. 38 of the Law No. 2016-1231 of for a Digital Republic added paragraph 10 to Art. L122-5 and paragraph 5 to Art. L342-3 of the Intellectual Property Code (Code de la propriété intellectuelle) (CPI); the Estonian law – Estonian Copyright Act, Art. 19(3). In 2021 German law was amended to incorporate the Directive and shifted from noncommercial TDM exception that it has enacted to years earlier to a full implementation of the Directive. Cf. Act on Copyright and Related Rights (Urheberrechtsgesetz, UrhG) as amended by Article 1 of the Act of 1 September 2017 (Federal Law Gazette I p. 3346); The Federal Ministry of Justice and Consumer Protection, translation by Ute Reusch, section 60d, available at [http://www.gesetze-im-internet.de/englisch\\_urhg](http://www.gesetze-im-internet.de/englisch_urhg). Other countries also incorporated the Directive’s TDM exceptions in full. See, e.g., the Dutch law The Copyright Directive Implementation Act, December 29, Stb. 2020, 558; Annemarie Beunen, *Copyright directive series - a closer look at the Netherlands*, europeana pro (august 3, 2022), available at <https://pro.europeana.eu/post/copyright-directive-series-a-closer-look-at-the-netherlands>.

<sup>114</sup> See *supra* note 101 and accompanying text.

<sup>115</sup> See Article 243 of the Copyright Act in Singapore (available at <https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/?ProvIds=P15-#pr243>).

## State of Israel Ministry of Justice

its plan to make Singapore a global AI ‘hub.’<sup>116</sup> The TDM exception in Singapore is quite robust and includes, *inter alia*, application in commercial contexts, a broad and open list of actions that are considered TDM and rights to prevail over contractual provisions that might frustrate TDM.<sup>117</sup>

In 2022, the *UK* expanded the TDM exception that it first legislated in 2014. The original TDM exception in the *UK*, as in many other European countries, was applied only to noncommercial scientific research.<sup>118</sup> After creating a national AI program in the *UK* (and perhaps not unrelated to it) *UK* copyright law has lifted the commerciality restriction, and now permits the use of copyrighted works for TDM in commercial settings as well.<sup>119</sup>

Other countries consider adopting ML-specific exemptions as well. For example, in 2019, in light of Canada’s aspiration to promote AI in its territory,<sup>120</sup> a parliamentary committee on copyright law recommended to legislate a copyright limitation for “informational analysis” or add “informational analysis” as a category in the fair dealing section.<sup>121</sup> *Canadian* law entails no fair use provision, but rather the similar mechanism of fair dealing.<sup>122</sup> As of this writing, Canadian case law has not yet considered the application of fair dealing to ML.<sup>123</sup> The committee recommendations are yet to be implemented.

A legal system that clearly avoids a specific statutory exception to ML is the *United States*. In fact, the interpretation in this Opinion, namely that the use of copyrighted materials for ML datasets comprises fair use, was laid down earlier in U.S. case law, from which Israel adopted the fair use doctrine.<sup>124</sup> This decision was made in *Authors Guild v.*

---

<sup>116</sup> See Singapore Government, National AI Strategy (available in: at <https://www.smartnation.gov.sg/initiatives/artificial-intelligence>) (vision: “By 2030, Singapore will be a leader in developing and deploying scalable, impactful AI solutions, in key sectors of high value and relevance to our citizens and businesses.”).

<sup>117</sup> See § 244 of the Singapore Copyright Act.

<sup>118</sup> See Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, No. 1372, adding Article 29A to the Copyright, Designs and Patents Act 1988. The Regulations came into force on 1 June 2014.

<sup>119</sup> See Article 29A of the UK Copyright Act.

<sup>120</sup> See Government of Canada, Pan-Canadian Artificial Intelligence Strategy, available at <https://ised-isde.canada.ca/site/ai-strategy/en> (last visited: 8.28.2022).

<sup>121</sup> See DAN RUIMY, HOUSE OF COMMONS, STATUTORY REVIEW OF THE COPYRIGHTS ACT: REPORT OF THE STANDING COMMITTEE ON INDUSTRY, SCIENCE AND TECHNOLOGY 85-87 (June 2019).

<sup>122</sup> See Canada Copyright Law, R.S.C. 1985, c C-42, art. 29. A broad interpretation of the ‘fair dealing’ doctrine over the years dimmed the differences between ‘fair dealing’ and the U.S. fair use doctrines. The leading judgment regarding the fair dealing criteria in Canada is *CCH Canadian Ltd. v. Law Society of Upper Canada* [2004] S.C.R. 339, 342 (Can.).

<sup>123</sup> See Brown, Byl & Grossman, *supra* note 54, at 159-161; Christopher Guly, *Canada Is Gathering Public Input on Copyright Implications of AI, Internet of Things*, CIGI (august 13, 2021) (“Given the absence of a clear rule to permit ML in Canadian copyright law (often called a text and data mining exception), our legal framework trails behind other countries that have reduced risks associated with using data sets in AI activities”) available at <https://www.cigionline.org/articles/%20anada-is-gathering-public-input-on-copyright-implications-of-ai-internet-of-things/>.

<sup>124</sup> See James Grimmelman, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 658 (2016); Benjamin L.W. Sobel, *Artificial Intelligence’s Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 96 (2017).

## State of Israel Ministry of Justice

*Google* (the “Google Books Case”) in which the court concluded that the Google books project, in which Google scanned to its servers millions of books for the purpose of preserving them and making them available to the public albeit with certain limitations, falls under the fair use doctrine.<sup>125</sup>

Over the years, this case law was interpreted broadly, to enable all uses of works whose audience is a computer and not a human. In the words of the scholar James Grimmelmann –

**“[Q]uietly, invisibly, almost by accident, copyright has concluded that reading by robots doesn’t count.”<sup>126</sup>**

The past years saw a certain decline of the broad interpretation that “reading by robots doesn’t count.” Specifically, two recent cases can be construed as dimming the de facto legal certainty that automatic operations have enjoyed. These recent cases refused applying the fair use doctrine to uses by computer systems. One case involved a searchable news aggregation platform, and the other concerned a service that allowed search, viewing and copying of protected broadcasts.<sup>127</sup>

It is doubtful whether these two cases will have any effect on the issue of ML datasets. These judgments do not deal with learning systems or AI, and the Google Books precedent may very well continue to apply uninterruptedly to the ML situation in which it was created. Indeed, the difference between ML enterprises and the new cases is substantial. These two cases did have human audience: the use that the system made of the works included presenting of a large part of the original works to end users, thus harming the original market of the rightsholders, who themselves offered access to these works. In contrast, the typical use of ML does not grant humans access to the works contained in the datasets. Yet, these cases may reduce the liberty that automatic systems have enjoyed, and affect at a certain stage the policy towards ML.

In conclusion, many countries around the world evidently concluded that without allowing ML to make some limited unauthorized use of copyrighted works, they will not be able to attain their goals in the field of AI. Consequently, countries that considered this issue instituted ML exceptions, whether by specific provisions or by interpretation to existing doctrines.

The clear advantage of specific provisions lies in enhancing legal certainty for both the AI and the creative industries. But this advantage is short-lived and might turn into a burden after not much time. The TDM exception provides a clear example. Up to approximately a decade ago, the issues that kept jurists busy in the field of automatic uses

---

<sup>125</sup> *Id. See also, Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 2014 U.S. App. LEXIS 10803, Copy. L. Rep. (CCH) P30,617, 42 Media L. Rep. 1898, 111 U.S.P.Q.2D (BNA) 1001, 2014 WL 2576342 (United States Court of Appeals for the Second Circuit June 10, 2014, Decided).

<sup>126</sup> *See Grimmelmann, supra* note 124, *id.*

<sup>127</sup> *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F. 3d 169 (2nd Cir. 2018); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 543-44 (S.D.N.Y. 2013). *See also Lemley & Casey, supra* note 60.



## State of Israel Ministry of Justice

of works revolved around data collection and analysis. This is the reason why the TDM concept has become—and is to this very day—the paradigm for exceptions in the field of ML as well. However, the TDM framework views the issue through the lenses of data collection, and is only partially applicable to the creation of dynamic datasets for autonomous learning by machines. The change of prism leaves a number of open questions in the context of contemporary ML technologies. For example, the TDM assumption is that while copyrighted works might be collected, the works themselves are immaterial to the objective of the operation: typically, the objective is the data contained in these works rather than these works themselves.<sup>128</sup> Based on this assumption, one of the TDM exceptions in the EU obliges the deletion of the works after the collection of the information, as discussed. But this demand makes very little sense for actual ML enterprises, which typically invest immense resources in creating the dataset and which require the dataset for vital future uses, including maintaining and updating their ML projects.<sup>129</sup> Indeed, some countries defined the TDM concept broadly and included in it up-to-date functions of learning systems. But the problem is structural. Even modern definitions are eventually based on technological capabilities that are currently known. The ability of specific provisions to predict the future is naturally limited, and the price paid for certainty is thus very high. Future technologies might not fall within the bounds of the statutory exceptions, and the more novel they are the less likely it is that they were foreseen by today’s legislators and drafted into the exception. In fact, a specific provision might inadvertently be technology-regressive, because it would produce a disincentive to adopt novel technologies that are not enumerated in the statutory safe harbor. For this reason, we posit that regulation of the issue through interpretation of existing doctrines as proposed in this Opinion is a more adequate tool for this issue. The next Part explicates the scope of this Opinion.

### E. THE SCOPE OF THE OPINION AND EXCEPTIONS

As discussed, the interpretation provided in this Opinion is that existing doctrines in copyright law protect the creation of ML datasets. Based on this interpretation, AI enterprises are entitled to include copyrighted materials in their dataset for the purpose of training AI systems, under the conditions we discussed. For example, the use of works of one author for the purpose of imitating the author may be excluded from the Opinion. This Part will clarify the scope of this Opinion and its limits in practical scenarios.

---

<sup>128</sup> See also Gervais, *supra* note 69, *id.*

<sup>129</sup> See also European Parliament, *supra* note 46, at 12 (“In sum, existing exceptions and limitations might not offer a stable legal framework to safely engage in TDM research projects and invest considerable resources”).

# State of Israel Ministry of Justice

## 1. DERIVATIVE USES OF THE DATASET

AI enterprises use the dataset they created even after the initial training stage and after the system is up, for the purpose of monitoring and improvements. Indeed, AI systems are not ‘fire-and-forget’ systems.<sup>130</sup> Without monitoring, repeated learning and constant improvement of performance, the system may develop in unexpected directions and produce anomalous results.<sup>131</sup>

It is therefore clarified that this Opinion applies also to later use of the dataset for the purpose of monitoring and improving the system, to the extent that the original use is within the scope of these defenses. Notably, the transient use doctrine is inapplicable to such cases, because the dataset is saved in the system and is not transient.

## 2. USING AN EXISTING DATASET TO TRAIN ANOTHER SYSTEM

Does the Opinion allow datasets to be used only for the enterprise that created them or for other enterprises as well? On the one hand, sharing datasets between enterprises can increase resource-efficiency and reduce waste, by eliminating the need to recreate a dataset that already exists. Clearly, there is no benefit for copyright owners from the duplication of the effort to create a dataset, because the new dataset creator will also enjoy the fair use protection. On the other hand, allowing firms to share datasets between them without limitation might incentivize business practices that will tilt the balance and that will result in disproportionate harm to the copyright owners. In light of this analysis, it is useful to distinguish between a number of situations.

### *A. Reuse of Dataset by the Same Enterprise or Use by Service Providers*

One scenario that clearly falls within the scope of this Opinion is when the enterprise that created the dataset intends to make another use of the dataset, whether or not the new task is related to the original system. We posit that such a use is non-infringing if the original use was non-infringing. Compelling enterprises to create the same dataset twice is inefficient (and unenforceable) and will not benefit copyright owners, who will

---

<sup>130</sup> See *supra* Part B.

<sup>131</sup> See, e.g., James Vincent, *Google ‘Fixed’ Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech*, VERGE (Jan 12, 2018) <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai> (reporting the well cited case where Google’s face-recognition system identified dark-skinned faces as gorilla faces).

## State of Israel Ministry of Justice

not receive payment also for the repeated creation of the dataset. Another scenario that is not problematic is the transfer of the dataset to an outsourcing company to perform services for the enterprise, such as organizing or labeling the dataset, retraining or algorithmic monitoring of the system, etc.

### *B. Infrastructure Projects*

Certain ML projects are tasked with the creation of a trained model that is intended to serve as infrastructure for other AI systems. A good example of such a project is the national NLP program. This program was conceived after the State identified a market failure in the domain of NLP in Hebrew and Arabic that stems, *inter alia*, from the relatively small number of Hebrew and Arabic (in the local dialect) speakers. This market failure has resulted in inferior NLP-based systems in Hebrew and Arabic. One of the primary objectives of the Government program is to develop a trained model according that academic, government and private enterprises will be able to build upon when developing specific NLP uses in Hebrew and Arabic.<sup>132</sup>

An infrastructure project such as this can best attain its objective if follow-up systems will be able to use the original dataset. Otherwise, each of the systems that rely on the trained model will be required to invest resources for the purpose of developing their own datasets (or developing the same dataset again) instead of focusing on their own specific task, producing waste, inefficiency and inferior products. At the same time, the harm caused to the copyright owners as a result of such derivative uses of the dataset is very limited, if any. Consequently, we posit that this Opinion applies also to the use of the dataset by other companies in derivative projects.

### *C. Sharing Datasets Between Unrelated Enterprises*

A more problematic case is transferring datasets between unrelated parties. There are two possible scenarios. The first is the transfer of the dataset between unrelated AI companies. We find this scenario less likely because fierce competition in the AI market limits such a practice. To the extent that such a practice is pursued, it is not automatically protected under this Opinion and will be examined ad-hoc according to the fair use criteria, in the manner that such analyses are carried out at present.

---

<sup>132</sup> See Government of Israel, Press Release – The establishment of an association of Natural Language Processing (NLP) technology companies in Hebrew and Arabic (22.09.2020).

## State of Israel Ministry of Justice

The more concerning scenario is that this Opinion will be construed, by mistake, as allowing a business model whose entire purpose is the creation of datasets that include copyrighted works and selling them to third parties. It is therefore clarified that such business models are not protected addressed under this Opinion. Using the dataset for the purpose of training a machine is not the same as a transacting in the dataset itself.<sup>133</sup>

This Opinion thus deems as permissible the transfer of the dataset, whether or not for consideration, to related entities, such as in infrastructure projects, or when providing services to an existing system. However, the Opinion does not address the matter of sharing of datasets between unrelated parties. Needless to say, a transaction in the dataset itself cannot be considered incidental use of the dataset or the works comprising it. Nor can such a lasting, valuable dataset enjoy the transient copying protection.

### 3. “SHARING” A DATASET

For the same reason that this Opinion affords no safe harbor for datasets sharing between unrelated parties, the Opinion does not a-priori cover the making of a dataset available to the public (even on ML-oriented websites).<sup>134</sup> The reason for this is twofold. First, as discussed, this Opinion does not intend to address whether transactions in datasets, either by selling them or by making them available to the public should be deemed fair use or not. Second, the entity that makes the work available has little if any control over the uses that the public can do with the dataset or its content. Consider an enterprise that trains systems to write summaries of academic articles. Making this enterprise’s dataset available to the public will allow search engines to refer to these platforms users who look for access to the articles themselves, and may infringe the rights of the authors of such articles who offer them in their own platforms. Therefore, while not all cases where a dataset is made available to the public are automatically infringing, this Opinion does not afford it a-priori exemption from infringement either. Such cases would be examined ad-hoc, in the same way that such analyses are conducted today.

An important question may arise regarding the secondary liability of an enterprise that properly transferred its dataset to a third party to infringing uses by that third party. A comprehensive discussion of this issue exceeds the scope of this Opinion. Briefly, this Opinion does not intend to bring about any changes in the doctrines of secondary liability for copyright infringement.<sup>135</sup>

---

<sup>133</sup> An original dataset might be protected as a work of compilation, as defined in § 4 of the Copyright Act.

<sup>134</sup> See, e.g., Github, <https://github.com>.

<sup>135</sup> See §§ 47, 48, 48A, 51 of the Copyright Act; CA 5977/07 *Hebrew University in Jerusalem v. Schocken Publishing House Ltd.*, (Nevo, 20.6.2011).

# State of Israel Ministry of Justice

## 4. THE PRODUCTS OF ML

At the end of the day, ML is nothing but a means to train systems to perform tasks. Eventually, the system produces a product—a prediction, identification, classification, reevaluation, or a new work that is based on the learning process. This sub-Part concentrates on generative AI. Examples of generative AI abound. One AI system creates poetry in the style of Nathan Alterman,<sup>136</sup> another system ‘paints’ like Van Gogh<sup>137</sup> or Rembrandt,<sup>138</sup> another system writes academic articles,<sup>139</sup> and another system performs any song in the style of any singer.<sup>140</sup>

Generative AI raises copyright questions that exceed the boundaries of this Opinion, such as copyrightability and ownership.<sup>141</sup> For the purpose of this Opinion, we clarify one point only: the defenses that were discussed above regarding the creation of datasets *do not* apply automatically also to the resulting product of the AI system. In other words, *a work will not be considered non-infringing merely because it was created by AI*. Liability may very well attach to a ML product that interferes with one of the exclusive rights defined in section 11 of the Copyright Act if no copyright limitation is available.<sup>142</sup>

## 5. CONTRACTUAL OR TECHNOLOGICAL HURDLES

This Opinion aims to strike the right balance between copyright and ML though interpretation of existing copyright doctrines. But what if the license under which the user obtains access to the materials inhibits the use that this Opinion otherwise permits?<sup>143</sup> For example, copyright owners or content hosting platforms may (and often do) include Terms

---

<sup>136</sup> See Altermanator, an automatic generator of the poems of Natan Alterman, <http://altermanator.herokuapp.com>.

<sup>137</sup> See AI Van Gogh, [https://ouchhh.tv/ai-van-gogh\\_immersive-data-painting#:~:text=STATEMENT%20%3E,All%20of%20Van%20Gogh's%20works%20created%20during%20his%20lifetime%20were,works%20were%20brought%20to%20life](https://ouchhh.tv/ai-van-gogh_immersive-data-painting#:~:text=STATEMENT%20%3E,All%20of%20Van%20Gogh's%20works%20created%20during%20his%20lifetime%20were,works%20were%20brought%20to%20life) (last visited: 28.8.2021).

<sup>138</sup> See, e.g., The Next Rembrandt, <https://www.nextrembrandt.com>. For another key example see Abraham Project, <https://abraham.ai>.

<sup>139</sup> See BoredHuman.com, [https://boredhumans.com/research\\_papers.php](https://boredhumans.com/research_papers.php).

<sup>140</sup> See Knight, *supra* note 28.

<sup>141</sup> See *supra* notes 16, 37.

<sup>142</sup> Despite the analytical distinction between the input of the system and its output, ‘Chinese walls’ between these stages are undesired because they would frustrate legal action against an enterprise for its AI product. Rigid ‘Chinese walls’ will allow the enterprise to evade liability based on the argument that the learning stage—where it was involved—is permitted, while the output stage was created by the system alone, without the enterprise’s involvement. The system itself is of course not a legal entity, and cannot be sued. It is thus crucial to view the system holistically at the time of lawsuits without rigid ‘Chinese Walls.’ See also *supra* note 37; Nadia Banteka, *Artificially Intelligent Persons*, 58 HOUS. L. REV. 537, 593 (2021).

<sup>143</sup> See Christophe Geiger, *The Answer to the Machine Should not be the Machine*, EIPR(4) 121 (2008); Elkin-Koren, *supra* note 72.

## State of Israel Ministry of Justice

of Use that prohibit data collection or disable it via Technological Protective Measures (TPM or DRM – Digital Rights Management). Such practices can render the safe-harbor portrayed in this Opinion merely theoretical.<sup>144</sup>

The question whether fair use is protected from contractual or technological override has been extensively debated around the world.<sup>145</sup> In the ML context, the European Parliament opined that it is necessary to state in legislation that the right to collect data for ML datasets would prevail over contractual provisions of right holders or the owners of platforms holding the data.<sup>146</sup>

Israeli case law has yet to address whether contracts can override copyright limitations, in particular fair use. The starting point for the analysis is the principle of freedom of contracts.<sup>147</sup> Yet, despite the centrality of this principle, contracting parties are not free to restrict protections of societal values that exceed their private interests.<sup>148</sup> The law may prevent the enforcement of contracts that hinder public and social interests, via *inter alia*, the Standard Form Contracts Act of 1982<sup>149</sup> and the contractual doctrines of good faith and public policy.<sup>150</sup>

Consider first contractual limitations on data collection that originate in a standard form contract, such as a website's Terms of Use. The Legal Counsel and Legislation

---

<sup>144</sup> See Lucie Guibault, *Copyright Limitations and Contracts: An Analysis of the Contractual Overridability of Limitations on Copyright*, The Hague, Netherlands: Kluwer Law International (2002). For the enterprise, a contractual claim will be preferable over exposure in copyright that comes with damages without proof of damage. See *supra* note 47.

<sup>145</sup> See also Elkin-Koren, *supra* note 49. An extensive discussion in the United States came after *ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996). This question is related to another question that exceeds the scope of this Opinion - whether copyright exceptions form users-right or merely defenses. Canada recognized users-rights in *Canadian Ltd. v. Law Soc'y of Upper Canada*, 2004 SCC 13, [2004] 1 S.C.R. 339 (Can.) and in a series of later judgments. In Israel, courts rejected this approach in *CA Premier League*, *supra* note 21 (paras. 17-18) (“...some think that it is possible to consider the uses permitted under the new law as rights of users, in the sense that they can also be used as a positive argument, and not just as an argument of defense. However, I do not accept this argument”). See also CRAIG JOYCE, WILLIAM PATRY, MARSHALL LEAFFER & PETER JASZI, *COPYRIGHT LAW* (4th ed.) 715, 1998 (referring to fair use as “privilege,” “affirmative defense,” and “limitation” in the same page).

<sup>146</sup> See European Parliament, *supra* note 46, at 12 (“In order to promote TDM research, the EU legislator should clarify that protection against contractual and technological override should also be always extended to TDM mining both materials protected and not protected by IPRs, including those made available in a dataset”). This clarification is required in the EU amid ruling that websites can protect content beyond copyright. CJEU, C-30/12, *Ryanair Ltd v. PR Aviation BN* (15 January 2015), ECLI:EU:C:2015:10.

<sup>147</sup> See GABRIELA SHALEV & EFI TZEMACH, *CONTRACT LAWS*, chapter 2 (4<sup>th</sup> Edition, 2019). See also *Elkin-Koren*, *supra* note 49, at 374-375; TAMIR AFORI, *THE COPYRIGHT ACT* 193-194 (2012).

<sup>148</sup> See Shalev & Tzemach, *id.*

<sup>149</sup> See Standard Form Contracts Act 5743-1982. See also VARDA LUSTHAUS & TANA SPANIC, *STANDARD FORM CONTRACTS* 42 (1994).

<sup>150</sup> *Id.*, at 526 (Discussing contract interpretation); GABRIELA SHALEV & YEHUDA ADAR, *CONTRACT LAWS – REMEDIES* 91 (2009) (explaining remedies for breach of contract). For a general explanation regarding the public policy doctrine in contract law see Shalev & Tzemach, *id.*, at 640 *et seq.*

## State of Israel Ministry of Justice

Department (Civil Law) issued an opinion on Terms of Use that hinder fair uses in 2016.<sup>151</sup> Undoubtedly, the ‘take it or leave it’ nature of Terms of Use ordinarily render it a standard form contract.<sup>152</sup> Under the Standard Form Contract Act, unfavorable conditions in Terms of Use can thus be invalidated in court.<sup>153</sup> A provision that categorically prohibits data collection may well be construed as an unfavorable provision in a standard form contract, because it prevents users from performing uses that Copyright Law deems fair.<sup>154</sup>

The issue is more complex when the contract is not a standard form contract. Contracting parties restrict their rights and privileges in freely negotiated contracts all the time, and restricting fair use in a contract may not be any different.<sup>155</sup> Yet, in many cases, restricting one’s fair use, even willingly, may impose negative externalizations or others.<sup>156</sup> If courts decide that contracts cannot override fair use, contractual provisions that do so can be invalidated under section 30 of the Contracts Act (General Part).<sup>157</sup> Even without deciding if contracts can override fair use, courts can apply ‘softer’ tools when interpreting contracts, which can prevent override of fair use and facilitate the creation of datasets, in appropriate cases.<sup>158</sup>

### E. CONCLUSION

This Opinion analyzed whether ML enterprises are entitled to include copyrighted materials in ML datasets. It concluded that datasets generally fall under the fair use doctrine, and sometimes perhaps also under the incidental use doctrine or the transient copying doctrine. The exception is nondiverse datasets, such as ones that are designed to mimic the style a single author.

This position aims to boost innovation and maintain Israel’s status as a worldwide leader in ML and AI, without inflicting actual harms to copyright owners. The fact that all around the world, legal systems conclude that copyright law should enable the creation of

---

<sup>151</sup> See Ministry of Justice, Legal Counsel and Legislation Department (Civil Law), *The Relationship Between Copyright Permitted and Standard Form Contracts*, 2016. Available at [https://www.gov.il/BlobFolder/generalpage/legal-opinions-01/he/legal-opinions\\_lo-cl-18-7-2016.pdf](https://www.gov.il/BlobFolder/generalpage/legal-opinions-01/he/legal-opinions_lo-cl-18-7-2016.pdf).

<sup>152</sup> See LCA 5860/16 *Facebook Inc. v. Ohad Ben-Hemo*, (31.05.2018).

<sup>153</sup> See § 19 of the Standard Contracts Act 1982.

<sup>154</sup> See *supra* note 152. See also Lusthaus & Spanic, *supra* note 150, at 83 (explaining that deviation from statutory provisions—albeit dispositive ones—can be considered an unfavorable contractual provision in a Standard Form Contract). Katherine J. Strandburg, *Free Fall: The Online Market’s Consumer Preference Disconnect*, 2013 U. CHI. LEGAL F. 95 (2013); Lital Helman, *Pay for Privacy (Privacy) Performance: Holding Social Network Executives Accountable for Breaches in Data Privacy Protection*, 84 BROOK. L. REV. 532, 537-38 (2019).

<sup>155</sup> See, e.g., Greenman, *supra* note 12, at 333-334.

<sup>156</sup> See, e.g., CA 156/82 *Lipkin v. Dor HaZahav Ltd.*, SCR 39(3), 85, 94 (1985); AH 22/82 *Yules House Ltd. v. Raviv Moshe*, SCR 43(1) 441, 463 (1989).

<sup>157</sup> See, e.g., Lipkin, *id.*, p. 95. Cf. CA 11/84 *Rabinowitz v. SHLAV*, SCR 40(4), 533, p. 547 (1986).

<sup>158</sup> See *supra* note 147 and accompanying text.

## State of Israel Ministry of Justice

ML datasets provides further support for our position. At the same time, special cases—some of which may be unforeseeable at the present time—may justify deviation from this conclusion and may shift the balance towards enhanced copyright protection. Notably, this Opinion does not aim to define the legal status of the AI *product* and its safe harbor is only extended to the ML process.

A question may arise as to the desirability of statutory amendments or specific regulations, under section 19(c) of the Act or otherwise, to effectuate an ML safe-harbor and enhance certainty. Indeed, several countries have pursued such a track. At this stage, we decided against legislative amendments. Legislation will clearly enhance certainty for all parties involved. But the price of such a path will be way too high. Until legislation procedures are completed (a process that can takes years), both technology and the market conditions may change dramatically, and interpretive tools will become necessary again. The same concern applies with respect to the enactment of regulations, albeit to a lesser degree.

Indeed, there is an inherent difficulty in advancing technology-specific legislation in dynamic areas such as AI, where legislative provisions and definitions rapidly become obsolete.<sup>159</sup> Worse yet, specific legislation might be technology regressive. A statutory safe harbor for ML methods that are known to the legislator at present may generate a powerful incentive for innovators to use the protected technologies and methods, and to continue using them even after novel ones will penetrate the market. This incentive may have a negative dynamic effect as well. Developers of ML technologies may be discouraged from developing revolutionary ML methods out of fear of legal action or fear that these methods will not be adopted in the market. Thus, while some technological contexts do call for legislative intervention, we believe that a ‘soft’ dynamic regulation of the matter via this Opinion would generate better incentives for the different industry players for the benefit of society. The possibility to advance legislation or enact regulations will be reconsidered if this Opinion and the case law that will follow will fail to create certainty in the market along the lines of this Opinion.

In conclusion, contrary to other legal systems, Israeli copyright law enables dynamic legal development without changes in legislation, via provisions such as the fair use doctrine. This ability allows dynamic regulation that promotes legal certainty on the one hand without rigid statutory criteria that may not survive the test of time.

---

<sup>159</sup> See Lital Helman, *Curated Innovation*, 49 AKRON L. REV. 695 (2016).