Chapter 12
AI, Consumers & Psychological Harm

Przemysław Pałka*

Abstract

The chapter addresses the notion of psychological harm inflicted upon consumers by AI systems. It demonstrates that the risk posed to consumers' mental health by AI systems is real and should be addressed, yet the approach taken by the EU in its AIA Proposal is suboptimal. The Regulation would either outlaw certain AI systems used by companies like Meta, Google, or Twitter or leave them unregulated. The chapter calls for a more nuanced approach. On the one hand, the law's understanding of psychological harm needs to be clarified, as the notion is much murkier than one would assume. On the other hand, instead of deploying a one-size-fits-all solution, policymakers should differentiate the strategies for combatting psychological harm, depending on the precise kind of risk to be managed.

12.1 Introduction

This chapter addresses the notion of psychological harm[1] potentially inflicted upon consumers by AI systems. It ponders what phenomena could be considered psychological harm, analyzes how AI systems could be causing them, and provides an overview of the legal strategies for combatting them.

There are two motivations behind this choice of subject. On the one hand, the European Union's Artificial Intelligence Act Proposal (AIA)[2] explicitly invokes the term when outlawing certain artificial intelligence practices.[3] The AIA, promising to be the first horizontal regulation of artificial intelligence in the world, would not only apply extraterritorially but may also serve as a blueprint for AI regulation in other jurisdictions. Hence, for good or bad, it is prudent to scrutinize the promises and pitfalls of its logic. Moreover, even though the chapter's assessment of the AIA approach to combatting psychological harm is rather negative,[4] the fact that the AIA has tried to confront the problem should be considered laudable. The AIA would bring the issue of the

* The research leading to these results has received funding from the National Science Centre, Poland, project no. 2022/45/B/HS5/01419, titled "Consumer Law and the Attention Economy."

[1] The notion of psychological harm neither has a stable meaning in law or any other discipline, nor is it a term of art. Moreover, it might not be the most fortunate formulation, yet the European Union has chosen to deploy it in the AIA Proposal.

[2] The Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels 21 April 2021, COM(2021) 206 final 2021/0106(COD) (hereinafter, AIA). Note that, at the time of writing of this chapter, a final text has not been agreed upon but two other versions exist: the Council's, of November 25, 2022, available at https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf, and the European Parliament's, of June 13, 2023, available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[3] See below, section 2.

[4] See below, section 4.

technology's impact on consumers' mental health to the fore, in a move that many scholars have sought.[5] Hence, the chapter opens by analyzing the AIA's approach to psychological harm.

On the other hand, a growing body of empirical research points to the links between various AI-powered products (social media, chatbots, video games) and adverse impacts on consumers' mental health.[6] Consequently, in all the jurisdictions considering mental health as a human right, or at least recognizing the consumers' interest in mental health protection, the question of what to do with AI systems causing psychological harm becomes a policy challenge. Hence, it makes sense to take a closer look at the link between AI-powered products, law, and mental health and analyze the available strategies of legal reform.

The chapter's argument proceeds in four steps. First, section 12.2 scrutinizes the place of psychological harm in the AIA overall. Second, section 12.3 shows that the meaning of psychological harm is not clear in law or any other discipline and ponders how it could be understood in AI governance and legal discourses. Third, section 12.4 demonstrates that the AIA's approach to addressing the psychological harm is suboptimal and would leave the enforcers with a tragic choice: either outlaw certain AI systems or leave them unregulated.[7] Finally, section 12.5 surveys various possible ways to address the problem in a more nuanced manner.

The overall aim in this chapter is not to solve the problem but rather to reveal its underappreciated complexity. Recognition of this complexity will hopefully lead to interdisciplinary dialogue of lawyers, AI experts, and mental health experts to find the best solutions.

12.2 Psychological Harm in AIA Proposal

The AIA invokes psychological harm in only one, albeit important place, namely Article 5, which states:

> 1. The following artificial intelligence practices shall be prohibited:
>
> (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or *psychological harm*;

[5] J. N. Rosenquist, F. M. S. Morton, and S. N. Weinstein, 'Addictive Technology and Its Implications for Antitrust Enforcement' (2021) 100 *North Carolina Law Review* 431; A. Zakon, 'Optimized for Addiction: Extending Product Liability Concepts to Defectively Designed Social Media Algorithms and Overcoming the Communications Decency Act' (2020) 2020 *Wisconsin Law Review* 1107; P. Pałka, 'Private Law and Cognitive Science' in B. Brożek, J. Hage, N. Vincent (eds.), Law and Mind: A Survey of Law and the Cognitive Sciences, (2021), pp. 217–48.

[6] M. G. Hunt, R. Marx, C. Lipson, and J. Young, 'No More FOMO: Limiting Social Media Decreases Loneliness and Depression' (2018) 37 *Journal of Social and Clinical Psychology* 751–68; J. Lambert, G. Barnstable, E. Minter, J. Cooper, and D. McEwan, 'Taking a One-Week Break from Social Media Improves Well-Being, Depression, and Anxiety: A Randomized Controlled Trial' (2022) 25 *Cyberpsychology, Behavior, and Social Networking* 287–93; A. M. Bean, R. K. L. Nielsen, A. J. van Rooij, and C. J. Ferguson, 'Video game addiction: The push to pathologize video games' (2017) 48 *Professional Psychology: Research and Practice* 378–89.

[7] Though the European Parliament's version contains one amendment slightly nuancing this approach, namely the recognition of recommender systems of some social media platforms as high-risk.

> (b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or *psychological harm*; (…).[8]

The sanction for violating these prohibitions can be significant. The AIA foresees an administrative fine of up to €30 million or 6% of the company's total worldwide annual turnover (not net income) for the preceding financial year, whichever is higher.[9] To put it into perspective: Google's annual revenue for 2022 was $279.8 billion[10] and Meta was $116.6 billion.[11] If the AIA applied in 2023 the maximum fine for Google would be $16.79 billion while Meta would have to pay almost $7 billion.

A careful reading of Article 5 shows that it does not prohibit marketing or using AI systems that deploy "subliminal techniques beyond a person's consciousness" or that exploit "any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability" in order to "materially distort a person's behaviour" *unless* it happens in a manner "that causes or is likely to cause that person or another person physical or psychological harm."[12] Considering this conditional structure, some commentators have expressed skepticism about the usefulness of this proposed provision. For example, BEUC (a federation of European consumer organizations) called for adding "economic harm" to the list of recognized harms to expand the scope of prohibited types of AI-driven behavioral manipulation.[13] Under the AIA, the manipulation of persons' behavior in the economic[14] or political sphere[15] would not, in principle, be prohibited unless such manipulation led to physical or psychological harm.

Consequently, if Article 5 of the AIA is adopted in its current form, the scope of the prohibitions will turn on the interpretation given to the vague notion of psychological harm. Surprisingly, the AIA does not define this term, as if its meaning was clear. In fact, there is no agreed-upon understanding of psychological harm in law (tort, criminal, product safety, or antitrust) or in other disciplines (psychology or psychiatry).

Notably, the conditional structure of Article 5 works both ways. Psychological harm – whatever we take it to mean – would not be outlawed *per se.* Instead, the AIA prohibits it only as

[8] AIA, art. 5.1 (emphasis added).
[9] AIA, art. 71.3.
[10] Statista, "Annual revenue of Google from 2002 to 2022," retrieved from: https://www.statista.com/statistics/266206/googles-annual-global-revenue/.
[11] Statista, "Annual revenue and net income generated by Meta Platforms from 2007 to 2022," retrieved from: https://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/
[12] AIA, art. 5.1.
[13] BEUC, „Regulating AI to Protect the Consumer: Position Paper on the AIA," October 7, 2021, available at: https://www.beuc.eu/position-papers/regulating-ai-protect-consumer, pp. 1-2.
[14] Scholars have been raising alarm about the prospect of algorithmic manipulation in the market sphere for a while now. See, e.g., R. Calo, 'Digital Market Manipulation' (2013) 82 *George Washington Law Review* 995–1051; E. Mik, 'The erosion of autonomy in online consumer transactions' (2016) 8 *Law, Innovation and Technology* 1–38; P. Hacker, 'Manipulation by algorithms. Exploring the triangle of unfair commercial practice, data protection, and privacy law' (2021) *European Law Journal* 1–34.
[15] J. Zittrain, 'Engineering an Election' (2013) 127 *Harvard Law Review Forum* 335; Z. Tufekci, 'Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency' (2015) 13 *Colorado Technology Law Journal* 203; D. Susser, B. Roessler, and H. Nissenbaum, 'Technology, autonomy, and manipulation' (2019) 8 *Internet Policy Review*.

a consequence of an AI system deploying "subliminal techniques beyond a person's consciousness"[16] or exploiting "vulnerabilities of a specific group of persons"[17] to "materially distort" a person's behavior. Again, depending on the meaning ascribed to the general concepts invoked in these provisions, the prohibition could be broader or more limited. However, when analyzing the AIA as a whole – even though it only mentions psychological harm in Article 5 – one could argue that the law would regulate adverse effects on mental health more horizontally.

The AIA is much broader in scope than the issue of psychological harm noted in Article 5,[18] and so to appreciate the more horizontal consequences of adopting a specific understanding of psychological harm it is important to understand the logic of the AIA as a whole. First and foremost, the AIA would structurally make up a part of the Union's product safety legislation.[19] It does not create private rights of action, does not affect *inter partes* liability rules, and only foresees the application of administrative fines as the means for enforcement. Consistent with the current trend in the European Union, the AIA adopts a "risk-based" approach to regulation.[20] It distinguishes between four levels of risk posed by AI systems – unacceptable, high-risk, low-risk, and minimal[21] – and imposes different rules depending on what category a given system falls under. The Regulation's core of 46 articles lays down the rules governing the marketing and use of high-risk AI systems,[22] with one article devoted to unacceptable risk,[23] one to low-risk,[24] and one to minimal-risk AI systems.[25]

The scope of application – the definition of an "AI system" – is controversial, with different formulations suggested by the Commission in the original AIA Proposal,[26] the Council's most

[16] AIA, art. 5.1.(a).
[17] AIA, art. 5.1.(b).
[18] Not least because the final text has not yet been agreed upon. Readers interested in a robust and critical reconstruction of the AIA's logic and contents might want to consult: M. Veale and F. Z. Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22 *Computer Law Review International* 97–112; M. Ebers, V. R. S. Hoch, F. Rosenkranz, H. Ruschemeier, and B. Steinrötter, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 *J* 589–603; P. Hacker, 'A legal framework for AI training data—from first principles to the Artificial Intelligence Act' (2021) 13 *Law, Innovation and Technology* 257–301; D. Svantesson, 'The European Union Artificial Intelligence Act: Potential implications for Australia' (2022) 47 *Alternative Law Journal* 4–9; R. J. Neuwirth, *The EU Artificial Intelligence Act: Regulating Subliminal AI Systems* (Routledge, 2023).
[19] Veale and Borgesius, 'Demystifying...', 97.
[20] G. D. Gregorio and P. Dunn, 'The European risk-based approaches: Connecting constitutional dots in the digital age' (2022) 59 *Common Market Law Review*.
[21] Veale and Borgesius, 'Demystifying...', 98.
[22] AIA, arts. 6-51.
[23] AIA, art. 5, a part of which has been discussed above.
[24] AIA, art. 52, applying to all "AI systems intended to interact with natural persons" other than high-risk systems and, effectively, requiring only that such systems disclose that they are AI, with some special transparency requirements for emotion recognition systems and systems producing the so-called "deep fakes." For the discussion of the former, see M. Durovic and J. Watson, 'Nothing to Be Happy about: Consumer Emotions and AI' (2021) 4 *J* 784–93 and of the latter, see; B. Chesney and D. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 *California Law Review* 1753.
[25] AIA, art. 69, suggesting voluntary codes of conduct for the AI systems considered neither high-risk nor low-risk.
[26] AIA, art. 3.(1), reading: "<<artificial intelligence system>> (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I [machine learning approaches (…), logic- and knowledge-based approaches (…), statistical approaches, Bayesian estimation, search and optimization methods] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with"

recent General Approach[27] and the Parliament's version.[28] Nevertheless, both versions define "AI systems" very broadly, as software or systems generating outputs based on machine learning or logic- and knowledge-based approaches. This means that a vast range of applications already in use, if their outputs are not entirely predictable to the developer, fall under the notion of an AI system and will be governed by the AIA. However, it is important to remember that the bulk of the requirements only apply to high-risk AI systems.

An AI system would be considered high-risk either if it made up a component of a product already subject to EU safety regulations and is required to undergo a third-party conformity assessment (medical devices, boats, toys, etc.) or if its use falls into one of the areas listed in Annex III.[29] The latter group includes: (1) biometric identification and categorization of natural persons, (2) management and operation of critical infrastructure, (3) education and vocational training, (4) employment, workers management and access to self-employment, (5) access to and enjoyment of essential private services and public services and benefits, (6) law enforcement, (7) migration, asylum and border control management, and (8) administration of justice and democratic processes.[30] Consequently, the AIA's robust requirements would apply to many AI systems identified as high-risk but not to all AI systems considered to be high-risk by scholars.[31] Notably, AI systems used for advertising and marketing, content moderation, price discrimination or search, or recommendations – despite potentially posing significant threats to consumers' interests[32] – would not be considered high-risk and, therefore, not subject to the vast majority of the AIA's rules. However, the Parliament's version would consider the recommender systems of very large online platforms as high-risk.

What would be the consequences of considering an AI system high-risk? Most importantly, a provider of such a system would have to establish a "risk management system"[33] accounting for, among other things, known and foreseeable risks associated with the AI system's use and adoption

[27] Council's version, art. 3.(1), reading: "<<artificial intelligence system>> (AI system) means a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts;"

[28] Parliaments' version, amendment 165, reading: "<<artificial intelligence system>> (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments;"

[29] AIA, art. 6.

[30] AIA, Annex III. Note that each of the areas contains further specifications, omitted here for the sake of brevity.

[31] I. Ajunwa, 'Algorithms at Work: Productivity Monitoring Applications and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law' (2018) 63 *Saint Louis University Law Journal* 21; R. Xenidis, 'Tuning EU equality law to algorithmic discrimination: Three pathways to resilience' (2020) 27 *Maastricht Journal of European and Comparative Law* 736–58; C. Longoni, A. Bonezzi, and C. K. Morewedge, 'Resistance to Medical Artificial Intelligence' (2019) 46 *Journal of Consumer Research* 629–50; A. G. Ferguson, 'Policing Predictive Policing' (2016) 94 *Washington University Law Review* 1109; S. Wilkinson, 'Artificial intelligence, facial recognition technology and data privacy' (2020) 3 *Journal of Data Protection & Privacy* 186–98.

[32] Mik, 'The erosion...'; R. A. Woodcock, 'Big Data, Price Discrimination, and Antitrust' (2016) 68 *Hastings Law Journal* 1371; J. M. Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2017) 51 *U.C. Davis Law Review* 1149; A. Jabłonowska, M. Kuziemski, A. M. Nowak, H.-W. Micklitz, P. Pałka, and G. Sartor, *Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence: Final Report of the ARTSY Project* (2018); L. Willis, 'Deception by Design' (2020) 34 *Harvard Journal of Law & Technology* 116–90; Zakon, 'Optimized for...'.

[33] The AIA uses the term "system" in both cases and so I do not use a different notion here, not to distort the meaning.

of suitable risk management measures.[34] These measures should include data quality assessment,[35] user-oriented transparency,[36] or human oversight.[37] It is, therefore, up to the provider (and, in some cases, standardization and certification bodies)[38] to decide what possible outputs of the AI system pose risks to health, safety, or fundamental rights.[39] The AIA does not define "risk," leaving it to others such as providers of AI systems (generally, private, for-profit entities).

Nevertheless, as the AIA often invokes the protection of fundamental rights,[40] and the EU's Charter of Fundamental Rights guarantees the right to mental integrity,[41] an argument can be made that threats to this right should be accounted for and mitigated in the case of high-risk AI systems. Violations of fundamental rights *per se* are not prohibited under the AIA but should be treated as risks requiring their acknowledgment and management. Consequently, unlike the practices mentioned in Article 5, the usage of AI systems mentioned in Annex III (in education, employment, or border control) causing psychological harm would not be prohibited. However, these risks would have to be accounted for and mitigated as any other risks to health and safety. The failure to take measures to prevent harm would be subject to a fine of up to €20 million or 4% of the company's total revenue for the preceding year, whichever is higher.[42]

Two observations follow from the analysis up to this point. First, even though the AIA only mentions psychological harm in Article 5, under a pro-consumer or systemic interpretation, the meaning of the notion is relevant to the much broader category of high-risk AI systems. In the case of Annex III systems, an implied obligation to minimize the harm should be recognized. Second, if one accepts this interpretation, the obligation to minimize psychological harm would only apply to high-risk AI systems. Thus, it would not cover a wide range of consumer-oriented online services like social media, video games, streaming platforms, or search engines, as these are considered low-risk.[43]

Consequently, regardless of whether the AIA is adopted in its current form, the meaning of psychological harm deserves scrutiny from scholars interested in AI regulation and governance. Under the AIA, the term would determine the scope of application of far-reaching prohibitions and requirements for self-regulation. In the unlikely event that the AIA is not enacted, the fact that AI systems might cause psychological harm will remain a policy challenge. The next section explores the meaning of psychological harm that can be used to guide policy decisions in this area.

12.3 What Could Psychological Harm Mean?

As mentioned in the previous section, the AIA does not define psychological harm. Moreover, the notion does not have a precise meaning in law or other disciplines. Hence, in this section, I want to survey various possible ways to understand the concept. I assume that the staffers preparing the AIA used the term psychological harm intuitively, as an umbrella category, to capture a wide range

[34] AIA, art. 9.
[35] AIA, art. 10.
[36] AIA, art. 13.
[37] AIA, art. 14.
[38] Veale and Borgesius, 'Demystifying...', 104–6.
[39] AIA, arts. 10, 13-14.
[40] *Id.*
[41] European Charter of Fundamental Rights, Art. 3., stating "Everyone has the right to respect for his or her physical and mental integrity."
[42] AIA, art. 71.4.
[43] As they are not listed in Annex III yet are designed to interact with natural persons.

of phenomena negatively affecting a person's mental condition. It will be up to the private parties, administrative agencies, and ultimately courts to give the notion concrete meaning. This is likely to be contentious debate.

Psychological harm is not a term of art in consumer law, or other areas of law. There is no agreed-upon definition and the term itself is used infrequently and inconsistently. Rachel Bayefsky, writing about tort law, noted that: "Neither courts nor scholarly commentators have engaged deeply with the question of what counts as psychological harm or how psychological harm relates to concepts such as expressive, emotional, or aesthetic harm, or the ideas of offense or insult."[44] Scholars analyzing psychological harm and the law have sometimes equated the notion with "serious emotional and mental conditions,"[45] psychological trauma,[46] emotional harm, and emotional distress.[47] Infliction of emotional distress, both intentional and negligent, is a contentious subject in tort[48] and criminal law.[49] Moreover, "emotional distress" is not synonymous with all possible instances of psychological harm, like addiction, eating disorders or a generalized anxiety disorder.

It is questionable whether it makes sense to try to establish what psychological harm *should* mean by analyzing the law as it is. The law may offer tangential insight in its historically limited ability to distinguish between "real" and fraudulent claims,[50] due to taboo,[51] or social prejudices.[52] On the whole, however, the law has treated psychological harm with less rigor or seriousness than our current sensibilities would demand.[53] Nevertheless, the law serves as a repository of wisdom about what psychological harm *could* mean, especially concerning the functioning of AI systems.

First, there is a question of whether to qualify certain adverse mental impacts as harm in their own right or only when they are consequences of an action already deemed unlawful. Bayefsky, when arguing that psychological harm should be recognized as a basis for standing in tort cases, suggests that "the harm must be a response to the alleged invasion of a legally protected interest (constitutional, statutory, or common law)."[54] Avlana Eisenberg documented and criticized the trend to criminalize "the infliction of emotional harm independent of any physical harm or threat of physical injury."[55] This raises the paramount question of whether we want to use the notion of

[44] R. Bayefsky, 'Psychological Harm and Constitutional Standing' (2016) 81 *Brooklyn Law Review* at 1564.
[45] I. Kilovaty, 'Psychological Data Breach Harms' (2021) 23 *North Carolina Journal of Law & Technology* 1–66 at 1.
[46] J. Dillard, 'A Slaughterhouse Nightmare: Psychological Harm Suffered by Slaughterhouse Employees and the Possibility of Redress through Legal Reform Note' (2008) 15 *Georgetown Journal on Poverty Law & Policy* 391–408.
[47] J. Ahuja, 'Liability for Psychological and Psychiatric Harm: The Road to Recovery' (2015) 23 *Medical Law Review* 27–52.
[48] D. Crump, 'Negligent Infliction of Emotional Distress: An Unlimited Claim, but Does It Really Exist' (2016) 49 *Texas Tech Law Review* 685–702; D. Crump, 'Rethinking Intentional Infliction of Emotional Distress Developments' (2017) 25 *George Mason Law Review* 287–301.
[49] A. K. Eisenberg, 'Criminal Infliction of Emotional Distress' (2015) 113 *Michigan Law Review* 607–62.
[50] B. J. Grey, 'The Future of Emotional Harm' (2014) 83 *Fordham Law Review* 2605–54.
[51] Ahuja, 'Liability for...', 29.
[52] J. Conaghan, 'Law, harm and redress: a feminist perspective' (2002) 22 *Legal Studies* 319–39 at 320.
[53] J. J. Kircher, 'The Four Faces of Tort Law: Liability for Emotional Harm' (2006) 90 *Marquette Law Review* 789–920 at 789; Ahuja, 'Liability for...' (writing: 'Judicial approaches to emotional harm claims are extensively criticised for confounding legal principle and defying logic, and there is general agreement that the English law of liability for causing mental injury is "in a dreadful mess".').
[54] Bayefsky, 'Psychological Harm...', 1560.
[55] Eisenberg, 'Criminal...', 609.

psychological harm to broaden the scope of conduct considered unlawful in our societies, or should we simply use it to increase the damages and fines for the already unlawful conduct?

When interpreting the AIA, the answer to this question could translate into very different legal tests and consequences. Should one opt for the former, then the test would be: does the law already consider an "action by" the AI system as illegal under tort, criminal, or any other body of law? If yes, then the result of the Article 5 prohibitions, or the need to manage the risk posed by high-risk AI systems, would merely be to add a new (though significant) sanction to be administered by a public body on top of preexisting civil or criminal liability. However, should one opt for the latter, the AIA could become a self-standing basis for considering specific mental impacts, until now lawful, as psychological harm. As we shall see in the following sections, there are policy reasons to consider the latter approach seriously.

Second, there is a question of what adverse mental impacts should fall into the scope of psychological harm. As the law currently does not give a definitive answer, various scholars have proposed their own working definitions. For example, Bayefsky defines psychological harm:

> as mental or emotional suffering or distress. This definition is intended to be broad and to encompass a wide variety of psychological reactions, including fear, sorrow, humiliation, and anger. (…) For example, the definition of psychological harm presented here includes a sense of offense or insult – harms that courts are especially wary of cognizing as injury-in-fact.[56]

Importantly, Bayefsky has constructed this definition, having previously proposed to limit actionable harm to the consequences of an action infringing upon a legally protected interest. However, even if one does not necessarily agree with this choice, the definition indicates a major challenge: psychological harm can refer to a wide range of occurrences. For example, if a consumer using an AI system – say, a chatbot – experiences fear, humiliation, or insult,[57] should that be considered psychological harm? For the purposes of the Article 5 prohibitions, this might be excessive, yet when it comes to risk-management systems for high-risk AI systems, this might be precisely the consequence a developer should try to avoid.

From the point of view of consumer law, one could argue for a more concrete threshold such as the protection of mental health. As one of AIA's goals is to guarantee consumers' health and safety,[58] in a time when mental health for overall individual and social welfare has risen in importance,[59] an option would be to gradually broaden the scope of consumer law's interest towards mental health and to equate psychological harm with risks to mental health.[60] Yet, the notion of mental health is not clear-cut either.

It is prudent to acknowledge that, contrary to normative disciplines like law or ethics, psychology and psychiatry are not primarily devoted to the study of harms. Psychology attempts

[56] Bayefsky, 'Psychological Harm...', 1565.
[57] For a vivid example of how this can happen, see: K. Roose, 'A Conversation With Bing's Chatbot Left Me Deeply Unsettled' (2023) retrieved from: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.
[58] G. Howells, C. Twigg-Flesner, and T. Wilhelmsson, *Rethinking EU Consumer Law* , 1st edition ed. (Routledge, 2017) pp. 15, 35.
[59] J. Ohrnberger, E. Fichera, and M. Sutton, 'The relationship between physical and mental health: A mediation analysis' (2017) 195 *Social Science & Medicine* 42–49.
[60] As advocated in Zakon, 'Optimized for...'.

"to describe, understand, and predict" the functioning of brains and minds,[61] whereas psychiatry is a branch of medicine devoted to diagnosing and treating mental disorders.[62] These disciplines can provide us with an understanding of mental health, but a judgment on what should be considered a harm involves additional normative considerations. The definition of mental health widely accepted by psychiatrists and psychologists is the one adopted by the WHO: "A state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community."[63] Consequently, good mental health cannot be equated with an absence of mental disorders[64] ("mental illness")[65] but must be understood as a positive state of ability to cope with stresses, work fruitfully, and contribute to one's community. In this sense, "risks to mental health" cannot be understood solely as the product of environmental factors that cause (or have caused) a mental disorder, but as all occurrences that increase the risk of developing a disorder, including diminishing a person's ability to cope with stress. Just like smoking cigarettes can be considered harmful before it leads to cancer, or consuming high-cholesterol food before it leads to atherosclerosis, mental health is at risk much earlier than when a psychiatrist diagnoses a condition calling for medical treatment. This analogy has profound consequences for using harm to mental health as a threshold for considering something a psychological harm.

Simply put, a large swath of AI-powered applications already in use can be deemed to be a risk to mental health under the WHO's definition. For example, a study by Hunt et al. demonstrated a causal relationship between one's use of social media and experiencing loneliness and depression.[66] Similarly, Lambert et al. showed that one week hiatus from social media decreases states of depression or anxiety.[67] Also, there are numerous studies that document the correlation between the use of social media and eating disorders,[68] sleeping disorders,[69] or even self-harm.[70]

Importantly for our considerations, social media deploys AI systems to decide what users see on their newsfeeds, what content to recommend or display,[71] and how to optimize for more user

[61] F. McManus and G. Butler, *Psychology: A Very Short Introduction* , 2nd edition ed. (Oxford University Press, 2014) p. 4. British spelling original.
[62] T. Burns, *Psychiatry: A Very Short Introduction* , 2nd edition ed. (Oxford University Press, 2019) p. 6.
[63] P. Fusar-Poli, G. Salazar de Pablo, A. De Micheli, D. H. Nieman, C. U. Correll, L. V. Kessing, A. Pfennig, A. Bechdolf, S. Borgwardt, C. Arango, and T. van Amelsvoort, 'What is good mental health? A scoping review' (2020) 31 *European Neuropsychopharmacology* 33–46 at 35.
[64] *Id.*
[65] Burns, *Psychiatry*, p. 6.
[66] Hunt, Marx, Lipson, and Young, 'No More FOMO'.
[67] Lambert, Barnstable, Minter, Cooper, and McEwan, 'Taking a One-Week Break from Social Media Improves Well-Being, Depression, and Anxiety'.
[68] S. A. McLean, S. J. Paxton, E. H. Wertheim, and J. Masters, 'Selfies and social media: relationships between self-image editing and photo-investment and body dissatisfaction and dietary restraint' (2015) 3 *Journal of Eating Disorders* O21.
[69] R. Alonzo, J. Hussain, S. Stranges, and K. K. Anderson, 'Interplay between social media use, sleep quality, and mental health in youth: A systematic review' (2021) 56 *Sleep Medicine Reviews* 101414.
[70] A. M. Memon, S. G. Sharma, S. S. Mohite, and S. Jain, 'The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature' (2018) 60 *Indian Journal of Psychiatry* 384–92.
[71] L. V. Bryant, 'The YouTube Algorithm and the Alt-Right Filter Bubble' (2020) 4 *Open Information Science* 85–90.

engagement.[72] This is a feature of the so-called "attention economy,"[73] where consumer data is analyzed by algorithms that later suggest what to display, advertisements and other content, on consumers' computers or smartphones.[74] This economy, the foundation for the business models of companies like Google, Meta, or Twitter,[75] creates negative external effects on consumers' mental health. Should these adverse effects be considered psychological harm for the purposes of the AIA? The next section examines this issue and concludes that a more nuanced approach is need than the one prosed by the AIA.

12.4. AIA's Unfortunate Choice

Let us imagine that the AIA becomes applicable in the form proposed by the European Commission. How would one then assess the AI systems posing a risk to consumers' mental health? What would need to change for Facebook, Google, or Twitter before deploying AI systems to extract consumers' attention? It is important to remember that algorithms for search, content delivery, and product and service recommendations are *not* considered high-risk AI systems under the AIA.[76] They neither make up safety components of products already regulated for safety nor belong to any of the areas mentioned in Annex III. Consequently, none of the obligations regarding risk management apply to the providers of these services. Based on the studies discussed above,[77] interacting with these systems comes with risks to consumers' mental health. There are two options available relating to the AIA. The AIA would either outlaw the use of AI systems curating content in social media's newsfeeds or leave them completely unregulated, demanding solely that they inform consumers of being AI systems.

Consider a hypothetical situation where an AI Agency in one of the EU's Member States investigates a social media company's use of an AI system for content display in individual newsfeeds. It finds that the system's goal is to maximize the amount of time users spend using the service. The agency also discovers that the AI system achieves its goal by mixing in content that the particular user is likely to consider funny, interesting or arousing to extend the user's interest and time on the service. It recognizes this as a form of subliminal messaging. Imagine that, according to the agency's investigation, the approach results in consumers expending much more time on the platform than they intended, increase use of the platform, and describing their

[72] Zakon, 'Optimized for...'; Rosenquist, Morton, and Weinstein, 'Addictive...'; P. Palka, 'The World of Fifty (Interoperable) Facebooks' (2020) 51 *Seton Hall Law Review* 1193.
[73] T. Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (Vintage Books, a division of Penguin Random House LLC, 2017); T. Wu, 'Blind Spot: The Attention Economy and the Law Symposium: Innovative Antitrust' (2018) 82 *Antitrust Law Journal* 771–806; J. Trzaskowski, 'Data-driven value extraction and human well-being under EU law' (2022) *Electronic Markets*.
[74] Palka, 'The World of Fifty (Interoperable) Facebooks'.
[75] T. Hwang, *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet* (FSG Originals, 2020).
[76] Though Amendment 740 in the Parliament's version nuances this slightly, proposing to designate as high-risk "AI systems intended to be used by social media platforms that have been designated as very large online platforms within the meaning of Article 33 of Regulation EU 2022/2065, in their recommender systems to recommend to the recipient of the service user-generated content available on the platform."
[77] Hunt, Marx, Lipson, and Young, 'No More FOMO'; Lambert, Barnstable, Minter, Cooper, and McEwan, 'Taking a One-Week Break from Social Media Improves Well-Being, Depression, and Anxiety'.

relationship with the platform as an addiction.[78] The agency concludes that consumers' interaction with such systems, based on the empirical data, poses a risk to their mental health.

The agency may check whether the conditions of Article 5 are met. Is the AI system deploying "subliminal techniques beyond a person's consciousness?" Maybe, since the AI system is mixing various kinds of content to influence the consumer's actions. Further, does the AI system operate this way "to materially distort a person's behaviour?" The answer is that the extended time spent on the system is due to behavioral manipulation by the system. Finally, can this lead to psychological harm? If one assumes that an increased chance of depression, anxiety, or loneliness[79] could result from excessive use of the system or service, the answer is in the affirmative. Consequently, if the answer to all three questions is "yes," the use by a social media platform of an AI system that manipulates the consumer in to extending its time on the platform would be *prohibited*. This could be considered excessive.

However, if a site or platform is found not to use subliminal techniques then the use of an AI system would not be prohibited under the AIA. There is a plausible argument that the encouragement of spending more time on Facebook or YouTube should not qualify as a "material distortion" of a person's behavior. In this case, Article 5 would not apply. For political, not legal, reasons, governments are hesitant to outlaw the business models of some of the world's largest companies.

But the risk to mental health persists! Simply because one concludes that certain AI practices are not, or should not, be outright prohibited, it does not follow that there is no problem. And yet, in our hypothetical scenario, once an agency determines that Article 5 does not apply, the next step would be to check if the AI system under scrutiny is considered high risk. If it is deemed a low-risk the only avenue of regulation is when the provider self-labels itself as an AI system. The unfortunate choice for the agency will be to prohibit the AI-enhanced systems of Facebook, Google, and Twitter or leave them unregulated.

This is just one example of the gaps in regulation presented by the AIA. For instance, many contemporary computer games deploy software that may or may not be considered as AI systems under the AIA. The rationale for coverage is that AI-enhanced videos games come with the risk of addiction.[80] Or, as recently demonstrated, consumers may choose to use free AI-powered chatbots, like Chat-GPT, in search of psychological therapy,[81] with the attendant risks of a wrong diagnosis or harmful therapeutic recommendations. Under the AIA, all such systems would have to be either directly outlawed or not regulated at all.

12.5 Future of AI, Psychological Harm, and Law

In this last section, I survey possible approaches to mitigating psychological harm caused to consumers by AI systems. This section explores the different paths that lawmakers could take and their relative strengths and weaknesses. However, there is one normative theme that underlies

[78] Y. Hou, D. Xiong, T. Jiang, L. Song, and Q. Wang, 'Social media addiction: Its impact, mediation, and intervention' (2019) 13 *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*.
[79] Hunt, Marx, Lipson, and Young, 'No More FOMO'.
[80] Bean, Nielsen, van Rooij, and Ferguson, 'Video game addiction'.
[81] 'Hidden Use of ChatGPT in Online Mental Health Counseling Raises Ethical Concerns' (January 2023). Retrieved from: https://www.psychiatrist.com/news/hidden-use-of-chatgpt-in-online-mental-health-counseling-raises-ethical-concerns/.

whichever path is to be undertaken—psychological harm is a complex and nuanced problem and there is no simple, one-size-fits-all solution.

First, one path would be to embrace the AIA's approach to risk management and simply broaden the catalog of high-risk AI systems. In the current version "emotion recognition systems" are considered low risk, unless used in areas such as education, employment, or border control enumerated is Annex III. The easiest way to fix this would be to add another area to Annex III relating to "consumer-facing systems for recommendations, content curation, and chat."

The strengths of this approach would include (i) embeddedness in the AIA's overall architecture, (ii) simplicity, and (iii) its broad applicability. However, these are also sources of potential weaknesses. Most importantly, the AIA delegates a significant number of decisions (normative in nature, though disguised as technical) to the providers, consulting firms, and standard-setting bodies. This approach should work like traditional safety mechanisms ensuring that a robotic arm does not injure workers or that critical infrastructure is safe from cyberattacks. However, when dealing with amorphic notions like psychological harm, the risks to be managed may remain unclear. Section 3 above noted that domain experts are not in agreement on what constitutes psychological harm. Leaving decisions on this to private parties with financial stakes in marketing the product creates an inherent conflict of interest. The alternative would be for lawmakers to address these normative questions directly.

Second, we should consider unpacking the term psychological harm into smaller, more concrete notions. As discussed above, consumer-facing AI systems can cause negative emotions (like fear or sadness), can adversely affect one's mental health (contributing to depression or anxiety), can be addictive, and give consumers harmful mental health advice (chatbot as a therapist). Arguably, these harmful effects require different normative evaluations and different policy solutions. For example, addictive technologies powered by AI systems, like social media or video games, present problems of a different magnitude than AI systems contributing to the development of eating disorders or self-harm in teenagers. Whereas the former can be seen as harmful, they could be managed through regulation short of prohibition, such as the sale of addictive products, like tobacco and alcohol, where regulations require mandatory warnings or age limits. The latter, on the other hand, requires substantive regulation of algorithms. However, the regulation would need to be more granulated than outright prohibition since chatbots being used as therapists may be beneficial overall. However, what to do precisely with each and every kind of possible harm is a question requiring both a normative judgment and expert knowledge in psychology, psychiatry, and public health.

Third, we should seriously consider what elements of the potential regulation should be AI-specific and those that should be treated more horizontally. Are the normative questions about the regulation of psychological harm better addressed in a venue like the AIA or in general tort law or consumer law? On the one hand, one could argue that such questions belong in the general areas of law, and AI governance instruments, like the AIA, should be treated as more targeted tools to deal with specific problems posed by AI. On the other hand, as shown in Section 3 above, tort law, has had a difficult time dealing with the issue of psychological harm. The current political and public perception of the need to regulate AI is an opportunity to focus on the issues presented by addictive technologies or technologies harmful to mental health.

12.6 Conclusion

This chapter has argued that the problem of psychological harm to consumers caused by AI systems should be understood more broadly in the context of potential harm to their mental health, which exposes the complexity of applying the term to the various uses of AI. The existing legal discourses deploy the notion of psychological harm broadly as an umbrella category for a myriad of possible effects. To operationalize it, we need not to only clarify the harms to be regulated but also a more nuanced approach than the one proposed by the AIA.

This chapter's negative assessment of the AIA's approach to psychological harm should not detract from the fact that the AIA makes a laudable attempt to addresses the problem. The chapter highlights the dilemma facing regulatory authorities applying the AIA to AI systems used by companies like Meta, Twitter, and Google of either outlawing their use or leavening them unregulated. Neither one of these options are optimal. Yet, the very need to clarify the meaning of psychological harm, paired with public policy maker's interest in regulating artificial intelligence, is a good starting point for addressing the problem of harm caused by AI-powered software to consumer mental health.