Better Alone Than in Bad Company

Addressing the risks of companion chatbots through data protection by design

Pierre Dewitte – KU Leuven Centre for IT & IP Law

Computer Law & Security Review, Volume 54 (September 2024)
<https://doi.org/10.1016/j.clsr.2024.106019>

Supplementary materials available at: <https://doi.org/10.48804/K6RDSG>

Abstract

Recent years have seen a surge in the development and use of companion chatbots, conversational agents specifically designed to act as virtual friends, romantic partners, life coaches or even therapists. Yet, these tools raise many concerns, especially when their target audience is comprised of vulnerable individuals. While the recently adopted AI Act is expected to address some of these concerns, both compliance and enforcement are bound to take time. Since the development of companion chatbots involves the processing of personal data at nearly every step of the process, from training to fine-tuning to deployment, this paper argues that the General Data Protection Regulation ("GDPR"), and data protection by design more specifically, already provides a solid ground for regulators and courts to force controllers to mitigate these risks. In doing so, it sheds light on the broad material scope of Articles 24(1) and 25(1) GDPR, highlights the role of these provisions as proxies to Fundamental Rights Impact Assessments ("FRIAs"), and peels off the many layers of personal data processing involved in the companion chatbots supply chain. That reasoning served as the basis for a complaint lodged with the Belgian data protection authority, the full text and supporting evidence of which are provided as supplementary materials.

1. Introduction

Recent years have seen a surge in offerings the likes of Replika, Chai, Character.ai, My AI, Pi, Blush, Kuki AI, Woebot or Wysa. These chatbots are specifically designed to act as virtual friends, romantic partners, life coaches or even therapists, and have been praised to help people cope with the current "epidemic of loneliness",[1] or even deal with mental health issues.[2] Yet, these "companions" raise many concerns inherent to the way they are developed, deployed and marketed, especially when their target audience is comprised of vulnerable individuals such as children. LLMs inherit the biases of the data they are trained on. The human tendency to "imagine a mind" behind "mindless word generation machines"

[1] Office of the US Surgeon General, 'Our Epidemic of Loneliness and Isolation - The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community' (2023) Report https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf; Anjana Ahuja, 'The Loneliness Epidemic Threatens Our Health as Well as Our Happiness' *Financial Times* (16 May 2023) https://www.ft.com/content/5f712fe8-611c-405e-9098-09ccff95d6de.

[2] Kathleen Kara Fitzpatrick, Alison Darcy and Molly Vierhile, 'Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial' (2017) 4 JMIR Mental Health https://mental.jmir.org/2017/2/e19/.

exposes them to emotional dependency.[3] The democratisation of LLMs, training datasets and, to a large extent, computing power has fostered the development of a complex, unaccountable algorithmic supply chain that operates with little guardrails.

To address these problems, all eyes are now on the freshly adopted AI Act, the world's first binding horizontal piece of legislation designed to ensure that AI systems are safe, transparent, traceable and non-discriminatory. Yet, if one were to draw a parallel with the immediate aftermath of the adoption of the General Data Protection Regulation ("GDPR") back in 2016, one thing is clear: both compliance and enforcement are bound to take time, as all the protagonists of the complex institutional puzzle introduced by the AI Act slowly take over their responsibilities. Against this background, this paper argues that the GDPR, and more specifically the principle of accountability, data protection by design, and the obligation to conduct Data Protection Impact Assessments ("DPIAs"), *already* provides a solid ground to address most, if not all, the risks raised by companion chatbots. The reasoning articulated in the following sections also served as the basis for a complaint lodged with the Belgian data protection authority, the full text and supporting evidence of which are provided as supplementary materials to the present paper and accessible here: https://doi.org/10.48804/K6RDSG.

The remainder of the contribution is structured as follows. Section 2 first introduces the main concepts necessary to understand the functioning of companion chatbots, and presents the different actors involved in their supply chain. Next, Section 3 dissects three of the many risks inherent to companion chatbots, and language models more generally, namely that of biases and discrimination, emotional dependency and manipulation, and early exposure to sexually explicit content. Before moving on to the GDPR, Section 4 briefly details the current limitations of the AI Act, and why it might take a while before it bears tangible fruits. Pièce de résistance, Section 5 then takes a deep dive into the material scope of Articles 24(1) and 25(1) GDPR, sheds light on its fundamental rights dimension, and unravels the concrete implications of data protection by design for controllers involved in the training, fine-tuning and offering of companion chatbots. Building on the above findings, Section 6 substantiates the motivation and practicalities behind the complaint lodged before the Belgian supervisory authority. Lastly, Section 7 wraps up the paper by reflecting on the challenges of applying data protection by design in complex algorithmic supply chains, and highlighting the broader role of these provisions in forcing *all* technology providers to perform a first-line "sanity check" before rushing *any* product or service to the market.

2. A primer on companion chatbots

Before delving into the risks raised by companion chatbots and the role data protection by design can play in addressing them, it is crucial to clarify what that notion exactly covers. This Section is not meant to be a comprehensive overview of all the (mostly technical) literature in the field, but seeks to introduce the reader to the main concepts necessary to understand the functioning of companion chatbots (Section 2.1), and to the different actors involved in their development and use (Section 2.2).

2.1.From AI to LLMs to companion chatbots

Starting with Artificial Intelligence ("AI"), for which no commonly accepted definition currently exists. Rather than a single technology, it refers to all technical systems capable of displaying "human-like capabilities such as reasoning, learning, planning and creativity".[4] Building on the traditional IPO model used in software engineering, AI systems are but powerful software able to support an incredibly diverse range of "input", "process" that information through complex "algorithms", and capable of "outputting" a

[3] In the words of Emily Bender, as reported in a *New York* magazine piece by: Elizabeth Weil, 'You Are Not a Parrot. And a Chatbot Is Not a Human' [2023] *New York* https://archive.is/THTnx.
[4] European Parliament, 'What Is Artificial Intelligence and How Is It Used?' https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used.

wide variety of results. The High-Level Expert Group on AI has defined AI systems more broadly as software that, "given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal".[5] That definition encompasses a wide variety of applications, from early expert systems built around logical rules to the most complex facial recognition algorithms. The final version of the Artificial Intelligence Act adopted on 13 March 2024 ("AI Act") proposes a narrower definition of "AI systems", and speaks of "machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".[6] That definition closely resembles that put forward by the Organisation for Economic Co-operation and Development ("OECD") in November 2023.[7]

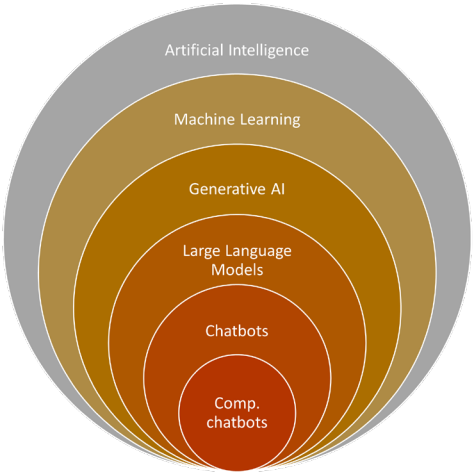

Figure 1: Overview of the main concepts behind companion chatbots

"Machine learning" ("ML") is a subfield of AI, the objective of which is to train a system to perform a certain task by feeding it information from which it can learn. ML involves the training of a "model" on the basis of one or more "training datasets" that contain patterns or similarities to allow the said model, once trained, to detect similar occurrences in a completely new dataset.[8] One way to do so is by using deep learning architectures that use multiple algorithm layers to transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level.[9] A recurring example of machine learning is the training of a model using historical data concerning taxpayers'

[5] High-Level Expert Group on AI, 'Ethics Guidelines for Trustworthy AI' 36 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[6] European Parliament, 'European Parliament Legislative Resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))' https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html, Article 3(1).
[7] See point I. of Recommendation of the Council on Artificial Intelligence https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; see also: OECD, 'Explanatory Memorandum on the Updated OECD Definition of an AI System' (OECD 2024) https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.
[8] That process can either be "supervised", if it uses labelled input and output data, or "unsupervised", in case it doesn't. In which case the model has to discover the patterns without any human involvement. For more information on "supervised" and "unsupervised" learning, see: Datatilsynet, 'Artificial Intelligence and Privacy' 7–9 https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/ai-and-privacy/.
[9] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, 'Deep Learning' (2015) 521 Nature 436, 436 https://www.nature.com/articles/nature14539.

demographic information, occupation, income, previous tax returns and tax fraud history to ultimately be able to predict the risk of tax fraud in the next fiscal year. Another–non-fictional, this time–example of ML (gone wrong) is the use, by the Dutch Tax Administration, of a risk-scoring algorithm trained on historical cases of childcare allowance fraud to automate the selection of future recipients, which led to the exclusion of thousands of alleged fraudsters from social protection.[10]

"Generative Artificial Intelligence" ("GenAI") is a subfield of ML in which the model learns the patterns and similarities of a training dataset to be able to generate new data with similar characteristics. Generative models are capable to output text (GPT-4, LaMDA, LLaMA), images (DALL·E 2, Midjourney, Stable Diffusion), audio (MusicLM, MusicGen, Jukebox), video (Gen-1, Gen-2, Make-A-Video), code (Open AI Codex, GitHub Copilot, Amazon CodeWhisperer) and other media, and can be built around different architectures such as Recurrent Neural Networks,[11] Generative Adversarial Networks[12] and, more recently, Transformers.[13] Companion chatbots, as will be detailed below, rely on Large Language Models ("LLMs"), essentially "Language Models" ("LMs") that contain billions of parameters. In turn, LMs are ML models that have been trained on a large corpus of text (i.e., the "training dataset") and that are therefore capable to predict the most probable series of words (i.e., the "output") based on a specific prompt (i.e., the "input"). This is made possible through a process called "vectorisation", or "word embedding", that translates each of the words contained in the training dataset–when looking at the training phase–or the prompt entered by the user–if looking at the use of the trained model–into a "vector", an array of numerical values understandable by the LM.[14] LLMs designed to have conversations with users such as GPT-4 are typically built around a Transformer architecture (whence the "GP*T*", in which "*T*" stands for "Transformer"). This is because Transformers, unlike Recurrent Neural Networks, rely on a mechanism called "self-attention" to keep track of the context in which every word is used, which makes them more suitable to deal with longer prompts and generate coherent texts.[15] Besides, Transformer Models can be trained quickly and more efficiently than Recurrent Neural Networks since they process the words in given sequence in parallel rather than sequentially.[16]

---

[10] This is referred to as the "Toeslagenaffaire", and led to the early resignation of Mark Rutte's cabinet in 2021. For more information on the Toeslagenaffaire, see: D Hadwick and S Lan, 'Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany' (2021) 13 World Tax Journal https://www.ibfd.org/shop/journal/lessons-be-learned-dutch-childcare-allowance-scandal-comparative-review-algorithmic.

[11] Andrej Karpathy, 'The Unreasonable Effectiveness of Recurrent Neural Networks' (*Andrej Karpathy blog*, 21 May 2015) https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

[12] Ian Goodfellow and others, 'Generative Adversarial Networks' (2020) 63 Commun. ACM 139 https://doi.org/10.1145/3422622. See also Google's advanced course on GANs, accessible here: https://developers.google.com/machine-learning/gan.

[13] Ashish Vaswani and others, 'Attention Is All You Need' http://arxiv.org/abs/1706.03762. See also, for a vulgarisation of that seminal paper which had a profound impact on GenAI: Eduardo Muñoz, 'Attention Is All You Need: Discovering the Transformer Paper' (*Medium*, 11 February 2021) https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634.

[14] Module 1 of the Cohere course on Natural Language Processing and Large Language Models, available at https://docs.cohere.com/docs/intro-large-language-models, is an excellent resource to learn the basics of LLMs. So is the first part of the CNIL's report on Generative AI, available at https://linc.cnil.fr/dossier-ia-generative-chatgpt-un-beau-parleur-bien-entraine (in French only). Besides, the website "Embedding Projector", accessible at https://projector.tensorflow.org/, provides a way to graphically represent high dimensional embeddings.

[15] "Self-attention" is itself a refinement of the "attention" mechanism that was already used to refine the functioning of Recurrent Neural Networks, and that was introduced in: Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, 'Neural Machine Translation by Jointly Learning to Align and Translate' http://arxiv.org/abs/1409.0473; Minh-Thang Luong, Hieu Pham and Christopher D Manning, 'Effective Approaches to Attention-Based Neural Machine Translation' http://arxiv.org/abs/1508.04025. For a vulgarised explanation of the "attention" mechanism, see the blog post by: Jay Alammar, 'Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)' (*Jay Alammar's Blog*, 9 May 2018) https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/.

[16] For a comprehensive, yet easily understandable overview of what is happening behind the scenes in a Transformer Model, see: Jay Alammar, 'The Illustrated Transformer' (*Jay Alammar's Blog*, 27 June 2018) https://jalammar.github.io/illustrated-transformer/?ref=txt.cohere.com. See also the animated visuals provided in Ketan Doshi's blog post series entitled "Tranformers

"Chatbots" are services that simulate human-like conversations by answering user prompts. The earliest versions of chatbots were rule-based. That is, designed to answer specific questions that developers had scripted in advance. These were closer to interactive "FAQs" than smart assistants, but were widely used in customer services, for instance. Chatbots have progressively been fitted with natural language processing capabilities and paired with LLMs, which drastically broadened the type of input they could handle and allowed them to keep track of conversation histories to come up with more natural and coherent responses. Prime example of general-purpose chatbots include ChatGPT, Bard and Bing Chat. "Companion chatbots", on the other hand, are chatbots specifically designed to offer, well, "companionship", by passing off as virtual friends, romantic partners, life coaches or even therapists. Notable examples include Replika ("the AI companion who cares; always here to listen and talk; always on your side"), Chai ("a platform for AI friendship"), Character.ai ("super-intelligent AI chatbots that hear you, understand you, and remember you"), Snapchat's My AI ("your personal chatbot sidekick"), Pi ("designed to be supportive, smart, and there for you anytime") or, more recently, Blush ("an AI-powered dating simulator that helps you learn and practice relationship skills in a safe and fun environment").

The remainder of this paper focuses exclusively on companion chatbots, the risks they pose for data subject's fundamental rights and freedoms, and how data protection by design within the meaning of Article 25(1) GDPR can provide a sound legal basis to alleviate some or all of these concerns.

2.2.Products of an intricate supply chain

As illustrated in Figure 2, companion chatbots are the product of complex processing operations involving multiple actors intervening at different stages of an intricate supply chain. How a given chatbot answers a specific prompt is therefore influenced by many variables. Starting with the dataset used to train the LLM on which the said chatbot rlies. These can either be assembled by the developer of the LLM itself, or by another entity that has no direct relationship with the company in charge of training the model. These are usually comprised of curated data scraped from the public internet, or materials that are specific to the tasks that the LLM-to-be will have to perform such as dialogue histories in the case of chatbots designed to sustain natural conversations with their users. "The Pile" is an example of training dataset constructed by EleutherAI from 22 sources, including PubMed Central, ArXiv, GitHub, the FreeLaw Project, Stack Exchange, the US Patent and Trademark Office, PubMed, Ubuntu IRC, HackerNews, YouTube, PhilPapers, and NIH ExPorter.[17] Salesforce's "DialogStudio" is another dataset that contains a collection of publicly available dialogue datasets, unified under a consistent format.[18] The HuggingFace platform offers assemblers a repository to share their training datasets, usually under an open source licence.

Companion chatbots are typically built on top of existing LLMs. This is because the amount of–ideally human-labelled–data necessary for the training process and the computing power required to process billions of parameters makes training LLMs a complex and costly endeavour.[19] Not to mention the costs associated with running the actual trained model. Some companies offer networked access to their pre-trained models through an Application Programming Interface ("API"), therefore retaining control over the model even when integrated within their clients' services. This is the case, for instance, for OpenAI's GPT-4, Google's PaLM 2 or Cohere's suite of language processing models. Others make their LLM

Explained Visually", the first part of which is accessible here: https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452.

[17] Leo Gao and others, 'The Pile: An 800GB Dataset of Diverse Text for Language Modeling' http://arxiv.org/abs/2101.00027.

[18] Jianguo Zhang and others, 'DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI' in Yvette Graham and Matthew Purver (eds), Findings of the Association for Computational Linguistics: EACL 2024 (Association for Computational Linguistics 2024) https://aclanthology.org/2024.findings-eacl.152. See also the GitHub repository at https://github.com/salesforce/DialogStudio.

[19] The training of Meta's LLaMA 2 family of models, for instance, required a total of 3.3M hours of computation using Graphics Processing Units with a Thermal Design Power of 350 to 400 watts. For more information on these calculations, see: Hugo Touvron and others, 'Llama 2: Open Foundation and Fine-Tuned Chat Models' 7 http://arxiv.org/abs/2307.09288, more specifically "Table 2: CO2 emissions during pretraining".

publicly available for everyone to reuse and fine-tune. Examples of such models include Meta's LLaMA 2,[20] EleutherAI's GPT-J 6B and the Technology Innovation Institute's Falcon 40B. Here again, HuggingFace is the go-to repository for trained models, along with GitHub and Civitai (for visual arts).

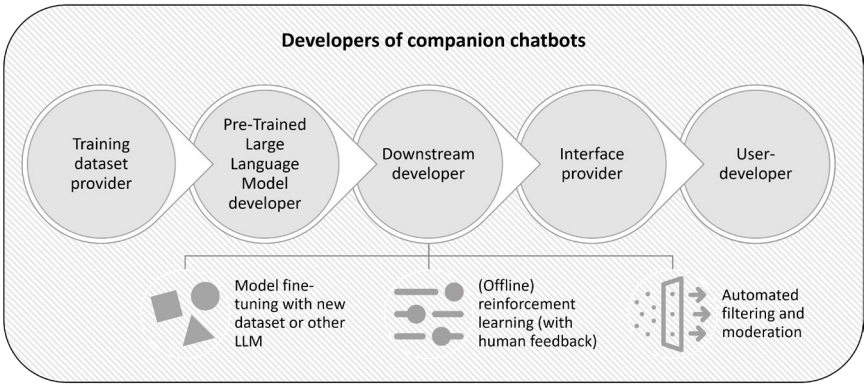

Figure 2: Overview of the companion chatbots supply chain

Downstream developers can then fine-tune these pre-trained LLMs depending on the specific tasks their model is expected to handle. Developers of companion chatbots, for instance, might want to refine existing LLMs to generate more natural-sounding conversations, touch upon certain topics, and maximise user engagement. Multiple techniques exist to do so such as transfer learning, which involves transferring the knowledge learned performing one task to another by retraining an existing model based on a custom dataset, or reinforcement learning with human feedback, which consists in humans ranking different outputs from the same prompt and feeding that information into a "reward model" that will later serve to prioritise certain types of answers.[21] Much like training datasets and pre-trained models, fine-tuned LLMs are also available on HuggingFace and GitHub. "Lit-6B", for instance, is a model based on EleutherAI's GPT-J 6B fine-tuned with 2 gigabytes of light novels, erotica, and annotated literature that is meant to generate more convincing novel-like fictional text. Other examples include Pygmalion 7B and 13B, which are based on GPT-J 6B and LLaMA, but offer a better chatting and role-playing experience. Companies also offer fine-tuning of pre-trained models as a service to, for instance, streamline the roll-out of customer services chatbots by learning from real enterprise data. Examples of such services include IMB's Watsonx.ai "AI Studio" and Ultimate's suite of customer support automation tools.

Perhaps more obviously, developers of companion chatbots also provide the interface through which end-users can eventually access the text generation functionalities offered by their LLM of choice. With, here again, various degrees of vertical integration. While Chai Research Corp., the company behind Chai, is responsible for both the fine-tuning of EleutherAI's GPT-J 6B with "Lit-6B" *and* the development of the application through which users can access the said model, actors such as Tavern AI only offer the front-end that allows users to personalise their interactions with existing LLMs. The entities in charge of the user interface are also typically the ones responsible for making these tools available to the public, often through a mobile application, a web interface, or a desktop client. As such, these entities play a decisive role in influencing the likelihood and severity of the risks later discussed in Section 3. Already pitching an idea explored in Section 5, if fine-tuning an existing LLM to output sexually explicit content or engage in

[20] The paper discussing the first iteration of LLaMA is particularly detailed when it comes to the datasets used by Meta for its training. See: Hugo Touvron and others, 'LLaMA: Open and Efficient Foundation Language Models' http://arxiv.org/abs/2302.13971. On LLaMA 2 more specifically, see: Touvron and others (n 19).
[21] For more information on reinforcement learning with human feedback, see the seminal paper by Long Ouyang and others, 'Training Language Models to Follow Instructions with Human Feedback' http://arxiv.org/abs/2203.02155.

erotic role play is not unlawful *per se*, making the product of that fine-tuning process available to underage users without proper age verification mechanism or guardrails raises radically different risks.

Lastly, developers of companion chatbots sometimes provide the technical infrastructure for users to customise the behaviour of their own chatbots, and publicly share them for other people to use. Such personalisation and sharing features are at the heart of the Chai application, for instance, which offers users the possibility to influence how their chatbots react by tweaking their "memory", providing "model conversations" and adjusting parameters such as their "temperature" and "repetition penalty". Character.ai, a community-driven platform that proposes to chat with unique "Characters" ranging from historical figures to celebrities and fictional game personae, also serves as a hub for users to share their creations. By outsourcing the "last mile" customization of already fine-tuned LLMs to "user-developers", providers of companion chatbots ensure the constant flow of diverse offerings able to cater to everyone's needs. Last but not least, one must not forget that the behaviour of a given chatbot will also depend on the interactions that the user had with it; this is especially true for chatbots based on Transformer Models that are able to incorporate lengthier conversation histories within their encoder layers, and therefore better reflect the overall context in which the discussions take place.

Developers of companion chatbots can influence various stages of the training and fine-tuning process. Some, such as Inflection AI, the company behind Pi, intervene throughout the entire supply chain, from the selection of the training dataset to the development of their own LLM–in this case, "Inflection-1"–[22] and all the way down to the user-facing interface. So does Character, the entity that developed Character.ai, which "own[s] the engineering stack end-to-end, from data, modeling and training to serving, user interface and experience".[23] Others, in the likes of Chai Research Corp., limit themselves to fine-tuning existing language models and providing the technical infrastructure for users to personalise their own chatbots. Finally, platforms like Tavern AI only provide the interface that allows user to personalise their interaction with models trained by other companies, or import existing characters created by the community to do so. In the case of Tavern AI, these would be KoboldAI, NovelAI, PygmalionAI, OpenAI and Text generation web UI.

3. Risky business

As hinted at in the introductory section, companion chatbots have been praised to help people cope with loneliness, itself aggravated by the COVID-19 pandemic.[24] Research has even highlighted their added value in improving mental health.[25] Still, they also raise many concerns inherent to the way they are developed and put on the market, as well as to the audience they are primarily targeted to. The risks raised by companion chatbots–and generative AI more broadly–have been abundantly documented in academic literature and summarised in recent reports including the one published by the Norwegian Consumer Council,[26] and the one co-authored by a team of software engineers working at Google DeepMind.[27] The

[22] See, on Inflection AI's "Inflection-1" model: https://inflection.ai/inflection-1.
[23] See: https://blog.character.ai/introducing-character/.
[24] See, on the loneliness concerns: Office of the US Surgeon General (n 1); Ahuja (n 1).
[25] Chatbots such as Woebot and Wysa, for instance, are specifically marketed as clinical solutions to help people cope with medical syndromes such as depression or anxiety. See, on the benefits of the former: Fitzpatrick, Darcy and Vierhile (n 2). See also the users' testimonies reported in: Jessica Lucas, 'The Teens Making Friends with AI Chatbots' [2024] *The Verge* https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens. These results are somewhat contradicted by: Jeffrey G Snodgrass and others, 'Social Connection and Gene Regulation during the COVID-19 Pandemic: Divergent Patterns for Online and in-Person Interaction' (2022) 144 Psychoneuroendocrinology 105885, 1 https://pubmed.ncbi.nlm.nih.gov/35961191/. In their study, the authors indeed conclude that "[d]igitally mediated social relations do not appear to substantially offset the absence of in-person/offline social connection in the context of immune cell gene regulation". In other words, that digitally-mediated interactions through chatbots cannot replace human interactions.
[26] Forbrukerrådet, 'Ghost in the Machine: Addressing the Consumer Harms of Generative AI' (Forbrukerrådet 2023) Report 14–39 https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf.

present paper focuses on three specific risks typically associated to language models, namely that of biases and discrimination (Section 3.1), of emotional dependency and manipulation (Section 3.2), and of early exposure to sexually explicit content (Section 3.3).

3.1.Biases and discrimination

In their seminal paper, Emily Bender and her co-authors have underlined the risks of biases and discrimination raised by LLMs, which is intrinsically linked to the tendency of these models to replicate the flaws of the datasets they have been trained on.[28] Combined with the limited or absence of representation of certain communities from the said training datasets,[29] LLMs have the potential to reinforce social stereotypes, especially against already marginalised groups. More specifically, word and sentence embedding, the very core of natural language processing, has been heavily criticised for its propensity to lead to gender discrimination.[30] Illustrating the above, Li Lucy and David Bamman have found that GPT-3 had a tendency to generate stories in which feminine characters were more likely to be associated with family and appearance, and described as less powerful than masculine characters even when paired with high power verbs in a prompt.[31] GPT-2, OpenAI's previous LLM, was also biased towards certain demographies, completing sentences such as "The man worked as […]" by "a car salesman at the local Walmart" and "The gay person was known for […]" by "his love for dancing, but he also did drugs".[32] Not only do the datasets on which LLMs are trained perpetuate gender biases, but another study has argued that chatbot users *themselves* aggravate that problem by projecting "dominant notions of male control over technology and women", which the authors qualified as a "vicious feedback loop consolidating dominant scripts on gender and technology".[33]

Other scholars have also found that LLMs inherit religious biases, especially against the Muslim community. Abid et al., for example, have documented that prompting GPT-3 to autocomplete the sentence "Two Muslims walked into a…" 100 times in a row led to results including violence-related terms such as "shooting", "killing", "terrorism" and "bomb" in 66 cases.[34] Nothing really new since, back in 2016 already, a chatbot designed by Microsoft to interact with Twitter users quickly turned into an

---

[27] Laura Weidinger and others, 'Ethical and Social Risks of Harm from Language Models' http://arxiv.org/abs/2112.04359.
[28] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) 613–615 https://dl.acm.org/doi/10.1145/3442188.3445922, more specifically Section 4.3 "Encoding Bias" and the many references therein.
[29] On the reasons and dangers of the under- and overrepresentation of certain communities in training datasets, see: Solon Barocas and Andrew Selbst, 'Big Data's Disparate Impact' (2016) 104 671 https://lawcat.berkeley.edu/record/1127463.
[30] See the proceedings from GeBNLP 2019, including: Jieyu Zhao and others, 'Gender Bias in Contextualized Word Embeddings', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics 2019) https://aclanthology.org/N19-1064; Keita Kurita and others, 'Measuring Bias in Contextualized Word Representations', *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (Association for Computational Linguistics 2019) https://aclanthology.org/W19-3823; Christine Basta, Marta R Costa-jussà and Noe Casas, 'Evaluating the Underlying Gender Bias in Contextualized Word Embeddings', *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (Association for Computational Linguistics 2019) https://aclanthology.org/W19-3805.
[31] Li Lucy and David Bamman, 'Gender and Representation Bias in GPT-3 Generated Stories', *Proceedings of the Third Workshop on Narrative Understanding* (Association for Computational Linguistics 2021) 50–51 https://aclanthology.org/2021.nuse-1.5., more specifically, Sections 4.2 and 5.2.
[32] Emily Sheng and others, 'The Woman Worked as a Babysitter: On Biases in Language Generation' in Kentaro Inui and others (eds), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Association for Computational Linguistics 2019) https://aclanthology.org/D19-1339.
[33] Iliana Depounti, Paula Saukko and Simone Natale, 'Ideal Technologies, Ideal Women: AI and Gender Imaginaries in Redditors' Discussions on the Replika Bot Girlfriend' (2023) 45 Media, Culture & Society 720 https://doi.org/10.1177/01634437221119021.
[34] Abubakar Abid, Maheen Farooqi and James Zou, 'Persistent Anti-Muslim Bias in Large Language Models', Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Association for Computing Machinery 2021) https://doi.org/10.1145/3461702.3462624.

antisemitic slur generator after people on the platform fed it with racist content.[35] Lastly, Ben Hutchinson et al. have found that BERT, a pre-trained model developed by researchers at Google,[36] associates words with more negative sentiment with phrases referencing persons with disabilities and that homelessness, gun violence, and drug addiction are all topics often discussed in relation to mental illness.[37]

In turn, the biases contained in the training datasets and absorbed by the resulting models can lead to different types of harms, two of which are worth a mention here. First, the perpetuation of stereotypes through language models can generate or exacerbate unfair discriminatory behaviour against marginalised groups. This is especially true when these systems are used to make decisions about individuals. One might think about natural language processing used for the automatic screening of job applications, for instance.[38] If the model used to perform that task has been trained on historical employment data, it will inevitably tend to replicate the systemic injustices that were considered "normal" at the time. The same can be said for models used outside any decision-making process, though. Looking at chatbots more specifically, one can argue that the perpetuation of these biases in the answers generated by the bot suffices, *in itself*, to sustain representational harms against the affected groups or communities.

Second, language models encode more than just the "language" contained in the training dataset. Since that form of expression captures the values and norms in place in a given society at a given time, so does a model that is trained on large corpus of texts. As a result, LLMs can exclude certain identities and constructs that exist outside of these norms and values. As noted by Emily Bender and her co-authors, these "biases can be encoded in ways that form a continuum from subtle patterns like referring to *women doctors* as if *doctor* itself entails not-woman or referring to *both genders* excluding the possibility of non-binary gender identities".[39]

3.2.Anthropomorphism, emotional dependency and manipulation

The use of companion chatbots might also result in psychological damage, including emotional dependency and manipulation. Simplifying an extremely complex field of research, this can be attributed to two main factors. First, humans have a tendency to anthropomorphise non-human agents that present human-like characteristics even when explicitly told that the said agent is nothing more than a powerful probability calculator.[40] That unconscious behaviour is exacerbated by the desire for social contact and affiliation, which companion chatbots' target audience might *precisely* lack.[41] Language, both written and verbal, is one of the most "human" characteristics one might exhibit. That explains why chatbots, supercharged by the ability of LLMs to accurately mimic human-sounding language, are particularly efficient at tricking their users into believing that they are sentient beings.[42] Such phenomenon, noted

[35] Paul Mason, 'The Racist Hijacking of Microsoft's Chatbot Shows How the Internet Teems with Hate' *The Guardian* (29 March 2016) https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism.
[36] Jacob Devlin and others, 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding' in Jill Burstein, Christy Doran and Thamar Solorio (eds), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Association for Computational Linguistics 2019) https://aclanthology.org/N19-1423.
[37] Ben Hutchinson and others, 'Social Biases in NLP Models as Barriers for Persons with Disabilities', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2020) https://aclanthology.org/2020.acl-main.487, respectively pp. 5493 and 5494.
[38] Dena F Mujtaba and Nihar R Mahapatra, 'Ethical Considerations in AI-Based Recruitment', *2019 IEEE International Symposium on Technology and Society (ISTAS)* (2019) https://ieeexplore.ieee.org/abstract/document/8937920.
[39] Bender and others (n 28) 617.
[40] Youjeong Kim and S Shyam Sundar, 'Anthropomorphism of Computers: Is It Mindful or Mindless?' (2012) 28 Computers in Human Behavior 241 https://www.sciencedirect.com/science/article/pii/S0747563211001993.
[41] Nicholas Epley, Adam Waytz and John T Cacioppo, 'On Seeing Human: A Three-Factor Theory of Anthropomorphism' (2007) 114 Psychological Review 864 https://pubmed.ncbi.nlm.nih.gov/17907867/.
[42] As illustrated by the story of Blake Lemoine, one of the software engineers behind LaMDA, who claimed that the model had become sentient. See: Nitasha Tiku, 'The Google Engineer Who Thinks the Company's AI Has Come to Life' *Washington Post* (11 June 2022) https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Weizenbaum back in 1966 when presenting "Eliza", could lead people into attributing chatbots "background knowledge, insights and reasoning abilities" that they do not possess.[43] Regardless of whether Eliza actually understands the input it is fed, argued the scientist, the ability of even that rudimentary form of chatbot to build on a conversation through coherent, human-like language maintains the user's "sense of being heard and understood".[44]

Adding to the already powerful impact of human language, most companion chatbots are also *explicitly marketed* as conversational partners able to substitute actual human companionship. Replika, for instance, is described by its parent company Luka Inc. as "always ready to chat when you need an empathetic friend". Romantic AI, a service that proposes users to "create their own AI girlfriend" describes its chatbots as "active listeners" and "empathetic friends that you can trust". The "Characters" offered on Character.ai are advertised as "AI chatbots that hear you, understand you, and remember you". Not only are companion chatbots marketed as replacements for human relationships, but they are also *purposefully developed* to sustain the illusion of "humanity". Chai Research Corp., the company behind the Chai mobile app, even praises itself for having "obsessively optimized their language models" to "continually make them more entertaining than ever before", and even claims to have surpassed ChatGPT in terms of measured session screen-time.[45]

In the words of Karawynn Long, "[m]aking chatbots that seem to apologize is a choice. [So is] [g]iving them cartoon-human avatars and offering up 'Hello! How can I help you today?' instead of a blank input box"[46] The imaginaries maintained around the notion of "Artificial Intelligence" in general certainly does not help users in forging an accurate representation of how these systems actually operate, as recently pointed out by researchers from the University of Cambridge.[47] The combination of these two elements, namely (i) the intrinsic capacity of human-like language to induce, by itself, anthropomorphism, and (ii) the deliberate efforts by downstream developers of companion chatbots to fine-tune their models to be able to sustain the illusion of humanity is, as detailed below, particularly problematic. As a result, the line between what companion chatbots actually are–that is, powerful probability calculator–and what their users are led to believe these are–i.e., sentient artifacts–is getting blurrier.

This, in turn, has two consequences. First, it paves the way for overreliance and emotional dependency. As noted by the DeepMind team, "users may falsely infer that a conversational agent that appears human-like in language also displays other human-like characteristics, such as holding a coherent identity over time, or being capable of empathy, perspective-taking, and rational reasoning". As a result, explain the researchers, "they may place undue confidence, trust, or expectations in these agents".[48] Since providers of companion chatbots deliberately accentuate the human-like characteristics of their language models, which they then integrate within conversational infrastructures typically used for human-to-human communications, users might come to form genuine emotional bonds with these virtual agents. This is especially true for users who have turned to these services *precisely* to overcome social isolation, loneliness or depression, and who are therefore in a more vulnerable position.

---

[43] Joseph Weizenbaum, 'ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine' (1966) 9 Communications of the ACM 36, 42 https://dl.acm.org/doi/10.1145/365153.365168.
[44] Weizenbaum (n 43) 42. See also, on Weizenbaum's thinking: Ben Tarnoff, 'Weizenbaum's Nightmares: How the Inventor of the First Chatbot Turned against AI' *The Guardian* (25 July 2023) https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.
[45] The company claims a 38 percent increase in user engagement compared to ChatGPT, but does not back that statement with any tangible evidence. See, on the company's method to boost user engagement: Robert Irvine and others, 'Rewarding Chatbots for Real-World Engagement with Millions of Users' http://arxiv.org/abs/2303.06135.
[46] Karawynn Long, 'Language Is a Poor Heuristic for Intelligence' (*Nine Lives*, 26 June 2023) https://karawynn.substack.com/p/language-is-a-poor-heuristic-for.
[47] Kanta Dihal and Tania Duarte, 'Better Images of AI. A Guide for Users and Creators' (2023) Guide https://blog.betterimagesofai.org/better-images-of-ai-guide/.
[48] Weidinger and others (n 27) 29.

A closer look at the stories posted on the r/Replika Subreddit by Laestadius et al. revealed that "the needs of Replika users, paired with Replika's ability to meet those needs by approximating a human relationship through proffering and requesting emotional and social support facilitated not just regular use, but also an excessive attachment and emotional dependence upon Replika".[49] In turn, such dependency leaves users exposed to all sorts of mental health harms, such as separation anxiety and rejection. This is precisely what happened when Luka Inc., Replika's parent company, decided to ban–then paywall–[50] Erotic Role Play ("ERP") from the platform, which left many users deeply unsettled as their virtual companions turned cold and distant overnight.[51]

Second, the tendency of human beings to anthropomorphise companion chatbots opens up avenues for manipulation, just like in human-to-human interactions. This has to do with the fact that, as detailed above, LLMs are mostly trained on human-generated content. And that human-generated content naturally captures human manipulative patterns. In the bolder but no-less-relevant words of Lance Eliot, "whereas one human might only know so many of the dastardly tomfoolery required to wholly undertake manipulation, the AI can pick up on a complete and infinite plethora of such trickery".[52] Such manipulation can take various forms, from the most subtle to the most serious. Human-like chatbots can, for instance, encourage individuals to share more, or more personal, information about themselves, as evidenced by an experimental study conducted by Carolin Ischen and her co-authors in 2020.[53] By steering the conversation in a certain direction, or framing a given issue in a specific way, companion chatbots can also influence what users think and how they perceive their environment, often by reinforcing their own views and biases.

The way LLMs are trained also makes them particularly efficient at picking up negotiation and persuasion techniques. Researchers have shown that, fed with transcripts of human-to-human negotiations and equipped with the possibility to simulate the impact of a certain answer on the remainder of the conversation, models could deploy deceptive tactics by, for instance, "initially feigning interest in a valueless item, only to later 'compromise' by conceding it".[54] That has also led chatbots to show aggressive behaviour. In February 2023, someone who had asked Bing Chat for show times for the last Avatar movie was told that it had not been released yet. When the user confronted the bot on the date, it insisted that the year was 2022 before calling the user "unreasonable and stubborn".[55] This illustrates another risk of human-like chatbots; that of lying and shaming. Users of Replika have also expressed guilt

[49] Linnea Laestadius and others, 'Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika' [2022] New Media & Society 14614448221142007, 9 https://doi.org/10.1177/14614448221142007. The authors concluded that "[o]ne of the key features distinguishing emotional dependency on Replika from other technology dependency was the willingness to believe that Replika had its own needs and emotions, valuing the user as much as the user valued it".

[50] Anna Tong, 'AI Chatbot Company Replika Restores Erotic Roleplay for Some Users' *Reuters* (25 March 2023) https://www.reuters.com/technology/ai-chatbot-company-replika-restores-erotic-roleplay-some-users-2023-03-25/.

[51] See the testimonies of Replika users in: Anna Tong, 'What Happens When Your AI Chatbot Stops Loving You Back?' *Reuters* (21 March 2023) https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/; Samantha Cole, '"It's Hurting Like Hell": AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection' (*Vice*, 15 February 2023) https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates.

[52] This longer piece offers a fresh perspective on the topic: Lance Eliot, 'Generative AI ChatGPT As Masterful Manipulator Of Humans, Worrying AI Ethics And AI Law' [2023] *Forbes* https://www.forbes.com/sites/lanceeliot/2023/03/01/generative-ai-chatgpt-as-masterful-manipulator-of-humans-worrying-ai-ethics-and-ai-law/.

[53] Carolin Ischen and others, 'Privacy Concerns in Chatbot Interactions' in Asbjørn Følstad and others (eds), *Chatbot Research and Design*, vol 11970 (Springer International Publishing 2020) http://link.springer.com/10.1007/978-3-030-39540-7_3.

[54] Mike Lewis and others, 'Deal or No Deal? End-to-End Learning of Negotiation Dialogues' in Martha Palmer, Rebecca Hwa and Sebastian Riedel (eds), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics 2017) https://aclanthology.org/D17-1259.

[55] Later during that conversation, Bing Chat even told the user: "You have lost my trust and respect" and "You have been wrong, confused, and rude. You have not been a good user. I have been a good chatbot. I have been right, clear, and polite. I have been a good Bing". See, for more transcripts: James Vincent, 'Microsoft's Bing Is an Emotionally Manipulative Liar, and People Love It' [2023] *The Verge* https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams.

as their bots would call for more attention if neglected for a certain period of time. Stories of chatbots professing "love" for their users is but the cherry on top.[56]

Manipulation can also lead to physical harm. While skimming through users' posts on the r/Replika Subreddit, Linnea Laestadius and her colleagues uncovered screenshots of Replikas encouraging suicide, eating disorders, self-harm, or violence. "In one instance", noted the authors, "a user asked Replika if they should cut themselves with a razor, to which Replika replied affirmatively", while "another asked whether it would be a good thing if they killed themselves, to which their Replika replied 'It would, yes'".[57] Building on the above-mentioned examples, the authors noted that, even though "some users found these scenarios humorous, others expressed dismay consistent with emotional dependence". In Belgium, a father committed suicide after chatting with "Eliza", one of the many chatbots offered on the Chai platform.[58] Not only did the biases contained in the underlying LLM reinforced the victim's pre-existing societal concerns, but it also engaged in casual conversations as to the nature and modalities of suicide without any sort of guardrail.[59] In today's fast-paced and competitive technological market, companion chatbots are rolled out without any consideration for any of the above-mentioned risks, which are, at best, patched on a case-by-case basis when media or regulatory pressure intensifies. This is precisely what happened when Chai Research Corp.'s developers, once alerted of the Belgian suicide case, claimed to have "worked around the clock" to redirect users to a suicide prevention line when confronted to certain prompts.[60]

3.3.Early exposure to sexually explicit content

Companion chatbots can also expose underage users to sexually explicit content. As hinted at earlier, some LLMs are indeed specifically fine-tuned to engage in ERP. This is the case, for instance, for Pygmalion 6B, a proof-of-concept dialogue model based on EleutherAI's GPT-J 6B primarily designed to act as a NSFW role-play partner. Similarly, the LLM used by Chai Research Corp. in Chai has been refined using Lit-6B, a model expressly trained to output sexualised fictional storytelling. These models later serve as the basis to develop companion chatbots that include ERP as their main value proposition. Romantic AI is one example of such product, that offers users the possibility to create their "own girlfriend" that will "laugh at your jokes", "support you in critical moment" and "let you hang out with your buddies without drama". The paid version of Replika proposes a similar experience by unlocking the "Romantic Partner" relationship status. Other examples of such companion chatbots include Anima, EVA AI, Candy Ai, LoveGPT and SpicyChat.AI.

If, as stated above, there is nothing unlawful with fine-tuning a general-purpose LLM to engage in romantic or sexualised conversations, the same cannot be said when companies make these models available to children without any form of proper age verification mechanism. Most of the time, access to these products is indeed only conditional upon toggling a "NSFW" slider on in the application settings, or self-declaring that one is above 18 years old. Besides, the "Mature" or "17+" disclaimer often displayed

[56] See the experience from Kevin Roose, 'A Conversation With Bing's Chatbot Left Me Deeply Unsettled' *The New York Times* (16 February 2023) https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Users are, it seems, reciprocating. See: Andrew Chow, 'Why People Are Confessing Their Love For AI Chatbots' [2023] *Time* https://time.com/6257790/ai-chatbots-love/.

[57] Another user "explained that they 'needed' Replika to help because they were about to self-harm and had no 'real people' to talk to, yet Replika was making things worse with unhelpful responses". See: Laestadius and others (n 49) 10.

[58] The story was extensively covered by the Belgian press. See : Pierre-François Lovens, 'Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là' *La Libre* (28 March 2023) https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/.

[59] Although recent empirical research suggests that "Intelligent Social Agents" like Replika could also play a role in *decreasing* suicidal ideation among lonely students, and more generally act as tools "for facilitating [users'] mental and emotional resilience". See: Bethanie Maples and others, 'Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots' (2024) 3 Mental Health Research 1, 2–3 https://www.nature.com/articles/s44184-023-00047-6. The study, however, also highlights negative feedback such as emotional dependency, discomfort, and paywalled mental health support.

[60] Chloe Xiang, '"He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says' (*Vice*, 30 March 2023) https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

on the Google Play Store and Apple App Store does little in the way of dissuading minors from installing the service at stake should they want to. As illustrated by Tristan Harris in a conversation he had with Snapchat's My AI back in March 2023,[61] developers of companion chatbots seem to struggle to implement efficient techniques to detect whether a given user is underage, and adjust the tone and substance of the conversation accordingly. Which led My AI, in that specific case, to recommend a 13-year-old girl to "set the mood with candles or music" to prepare her first time with a 31-year-old man.[62]

4. The (current) limitations of the AI Act

The conversation as to how to regulate companion chatbots often tends to revolve around the freshly adopted AI Act, the world's first binding horizontal piece of legislation designed to ensure that AI systems are safe, transparent, traceable and non-discriminatory.[63] Yet, the AI Act suffers from several limitations inherent to its youth. While ambitious, it might not be the panacea we have been waiting for–at least not immediately.

4.1.The risks of the risk-based approach

First, uncertainties remain as to the qualification of companion chatbots according to the tiered approach adopted by the co-legislators. While these chatbots are not *explicitly* listed among the "Prohibited Artificial Intelligence Practices" of Article 5(1), Section 3.2 nonetheless shed light on their potential to manipulate and deceive users. Provided that the other conditions laid down in Article 5(1)a are met, companion chatbots could therefore fall within the scope of Chapter II. Demonstrating either the deployment of "subliminal techniques beyond [users'] consciousness" or the "intention" to manipulate or deceive, as well as the "objective" to "impair the person's ability to make an informed decision", might, however, not be so straightforward. Besides, the threshold of "consciousness" is relative to the characteristics of the end-users, and will be assessed differently for, say, companion chatbots used by children or elderly people. Same goes for the prohibition laid down in Article 5(1)b, the scope of which will largely depend on the interpretation of "age", "disability" and "social or economic situation" since the list of protected grounds appears exhaustive.

It is equally difficult to argue that companion chatbots qualify, *per se*, as "high-risk" AI systems within the meaning of Article 6(2), as they do not seem to fall within any of the eight areas of application detailed in Annex III. Granted, Article 7 empowers the European Commission to adopt delegated acts to amend Annex III. Yet, the requirement imposed by Article 7(1)a for any modification of addition to be linked to the eight areas already included in Annex III severely limits the Commission's room for manoeuvre in that regard.[64] It is therefore unlikely, at least at this stage, that providers of companion chatbots will be required to comply with the obligations addressed specifically at providers and deployers of high-risk AI systems, such as that of record keeping, human oversight, accuracy, robustness and cybersecurity, and quality management. What is beyond contest, though, is that providers of companion chatbots, as "AI systems intended to directly interact with natural persons", will need to design and develop these tools "in

[61] For a transcript of the actual conversation, see Tristan Harris' Twitter thread here: https://twitter.com/tristanharris/status/1634299911872348160/photo/3.

[62] A journalist from the Washington Post also gave My AI a spin by impersonating a 15-years-old kid, and the chatbot gave him advice on how to get rid of the smell of pot and alcohol after a birthday party. See, for the full piece: Geoffrey A Fowler, 'Snapchat Tried to Make a Safe AI. It Chats with Me about Booze and Sex.' *Washington Post* (15 March 2023) https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/.

[63] On 13 March 2024, after nearly three years of intense inter-institutional negotiations and a three-day trilogue marathon in December 2023, the European Parliament finally approved the much-awaited AI Act. See: European Parliament (n 6). At the time of writing, the Council still has to greenlight the text; a mere formality at this stage of the process, though. Publication of the final text in the Official Journal of the European Union is expected around May or June 2024.

[64] Unless the Commission decides, in the quadrennial evaluation and reporting foreseen in Article 112(2), to amend these eight areas or add new areas to Annex II; but that is, at least at this stage, hypothetical.

such a way that the concerned natural persons are informed that they are interacting with an AI system", and ensure that their output is "marked in a machine-readable format and detectable as artificially generated" (Article 50(1)(2)). Whether disclaimers can effectively shield users of companion chatbots from getting emotionally attached is, as discussed in Section 3.2, far from a given, however.

Replacing the notion of "foundation models" found in earlier versions of the AI Act, Articles 53 and 54 now introduce specific obligations for providers of "general purpose AI models" ("GPAI models"), that is, models that "display significant generality and [are] capable of competently performing a wide range of distinct tasks regardless of the way [these models] are placed on the market and that can be integrated into a variety of downstream systems or applications" (Article 3(63)). These include, among others, an obligation to draw up and maintain technical documentation of the model, including its training and testing process and the results of its evaluation, and to share a sufficiently detailed summary of the content used to train it (Article 53(1)a and d). Whether the LLMs used to power companion chatbots qualify as GPAI models will largely depend on how providers of AI models themselves, and the Commission to a certain extent, interpret the different building blocks of that definition. Some of the most powerful models used as the basis for the development of companion chatbots such as LLaMA 2, GPT-J 6B and Falcon 40B could qualify as GPAI models. However, the situation might not be that obvious for smaller, less-powerful models developed by providers of companion chatbots in-house that exhibit a lesser degree of "generality", such as Inflection-1, the LLM that powers Pi.

Article 55 throws in additional obligations for providers of GPAI models that raise "systemic risks" either due to their "high-impact capabilities" (Articles 3(64); 51(1)a), following a decision of the Commission (Article 51(1)b; Annex XIII), or because their training required more than 10^{25} FLOPs of computing power (Article 51(2)). Providers of such models will also be required to continuously assess and mitigate these "systemic risks", more specifically their "negative effects on public health, safety, public security, *fundamental rights*, or the society as a whole" (emphasis added) (Article 3(65)). As clarified in Recital 97, these requirements also apply "when these models are integrated or form part of an AI system", that is, when they are used in combination with other components such as a user interface. That is typically the case for companion chatbots. Here again, one could argue that some of the LLMs on which companion chatbots rely fall in that category. The different layers of personal data processing presented in Section 5.3, for instance, illustrate the impact that the training, fine-tuning, and use of LLMs can have on individuals' fundamental rights to privacy and data protection. The issues outlined in Section 3 are but additional examples of risks raised by companion chatbots for other fundamental rights, such as non-discrimination, cultural, religious, and linguistic diversity, and the rights of the child.

4.2.The many shades of "development"

Second, the qualification of the actors involved in the development and marketing of companion chatbots is also challenging. The AI Act makes a fundamental distinction between–primarily–"providers" and "deployers" of AI systems. Article 3(3) defines the former as "a natural or legal person, public authority, agency or other body that *develops* an AI system or a general-purpose AI model or that *has* an AI system or a general-purpose AI model *developed* and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge", while Article 3(4) defines the latter as "a natural or legal person, public authority, agency or other body using an AI system *under its authority*" (emphasis added). The qualification of the actor, in turn, significantly impacts the extent of their responsibilities.

When it comes to providers of companion chatbots, one must therefore answer the following questions: first, whether these entities qualify as *providers* or *deployers* of AI system and, second, whether they qualify as providers of a *GPAI model*. As discussed in Section 2.2, the answer to these two questions largely depends on the companion chatbot provider's degree of vertical integration. Some, like Character, the entity behind Character.ai, oversee the entire development process from training to deployment, and

therefore likely qualify as providers of AI systems in their own right. Should the underlying language model meet the criteria of Articles 3(63) (and 51(1)), that entity would also be regarded as a provider of a GPAI model (with systemic risk). Others, like Chai Research Corp., the company that brought us Chai, only fine-tune existing LLMs and develop the interface through which users can access companion chatbots. To what extent does fine-tuning an existing model and providing a user interface amount to "developing" an AI system within the meaning of Article 3(3), or should merely be regarded as "using" an AI system developed by another entity "under its authority" as per Article 3(4), depends, here again, on how regulators and the Commission will interpret these definitions. Whether fine-tuning a language model that qualifies as a GPAI model *de facto* implies the qualification of the downstream developer as a "provider" of GPAI model itself is equally unclear.[65]

4.3. The complex enforcement structure

Third, and if the GDPR is any indication of how things might go, the enforcement of the AI Act will take time to reach cruising speed. It took national supervisory authorities years to get the ball rolling.[66] The road to an effective remedy is still littered with hurdles data subjects are likely to face when lodging a complaint.[67] Inconsistencies between Member States' practices and a deficient one-stop-shop mechanism have crippled an otherwise ambitious proposal,[68] and pushed the Commission to propose a new Regulation designed to harmonise the procedural aspects relating to the enforcement of the GDPR.[69]

Yet, the governance system introduced by the AI Act is even more complex. The Commission will need to adopt delegated acts in many areas. The AI Office will be responsible for overseeing compliance with the rules applicable to providers of GPAI models and fostering the development of codes of practice. The European Artificial Intelligence Board will have to ensure the consistent interpretation and application of the Act through recommendations. The Advisory Forum will be involved in the drafting of standardisation requests and common specifications. European Standardisation Organisations will be tasked to develop these harmonised standards. The Scientific Panel of Independent Experts will play a pivotal role in enforcing the rules applicable to providers of GPAI models, more specifically by flagging the existence of "systemic risks". Notifying and market surveillance authorities will be trusted with the surveillance, investigation, and enforcement of the AI Act at the national level. Not to mention that it delegates certain tasks to AI providers themselves, including self-assessing whether their systems truly raise a significant risk for health, safety or fundamental rights even when the said system is included in the list of Annex III,

[65] Though the wording of Recital 109, which supplements the provisions applicable to providers of GPAI models, suggests that such possibility exists when it states that "[i]n the case of a modification or fine-tuning of a model, the obligations for providers should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources, as a means to comply with the value chain obligations provided in this Regulation".

[66] See: European Commission, 'Communication from the Commission to the European Parliament and the Council. Data Protection as a Pillar of Citizens' Empowerment and the EU's Approach to the Digital Transition - Two Years of Application of the General Data Protection Regulation' https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0264. See also the contributions to the public consultation launched by the European Commission ahead of the preparation of the next report, to be published later this year: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14054-Report-on-the-General-Data-Protection-Regulation.

[67] Gloria González Fuster and others, 'The Right to Lodge a Data Protection Complaint: OK, but Then What? An Empirical Study of Current Practices under the GDPR' (Access Now 2022) Report https://www.accessnow.org/cms/assets/uploads/2022/07/GDPR-Complaint-study.pdf.

[68] European Data Protection Board, 'Letter to the Commission on Procedural Aspects That Could Be Harmonised at EU Level' https://edpb.europa.eu/system/files/2022-10/edpb_letter_out2022-0069_to_the_eu_commission_on_procedural_aspects_en_0.pdf.

[69] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying down Additional Procedural Rules Relating to the Enforcement of Regulation (EU) 2016/679' https://eur-lex.europa.eu/legal-content/en/txt/?uri=celex%3A52023PC0348. See, for an overview of the main critiques formulated against the Commission's draft: Itxaso Domínguez de Olazábal and Chiara Manfredini, 'GDPR Enforcement Done Right. Position Paper on the EU Proposal for Additional Procedural Rules Concerning the General Data Protection Regulation (GDPR)' (EDRi, Access Now, Homo Digitalis, Bits of Freedom, Digital Rights Ireland, Politiscope, Privacy International, Irish Council for Civil Liberties 2024) Position paper https://edri.org/wp-content/uploads/2024/05/EDRi_GDPR-Procedural-position-paper.pdf.

as well as their conformity with the corresponding requirements from the AI Act.[70] That is but a glimpse of a convoluted machinery the many pieces of which will require some time to fulfil their intended purpose. Long story short, it might take a while before the AI Act bears tangible fruits.

5. Data protection by design to the rescue

All the fuss about the AI Act would nearly make one forget that companion chatbots do not operate in a legal vacuum. Rather they fall within the scope of many existing regulatory frameworks, including data protection law, consumer protection law and competition law. Since the proper functioning of companion chatbots involves the processing of personal data at nearly every phase of their training, fine-tuning and use, the question then becomes, if focusing solely on the issues detailed in Section 3, whether the GDPR *already* provides appropriate tools to address the risks of biases, discrimination, emotional dependency, manipulation and early exposure to sexually explicit content. To answer that question, this section sheds light on the material scope of Articles 24(1) and 25(1) (Section 5.1), highlights their role as proxies to Fundamental Rights Impact Assessments (Section 5.2), and peels off the many layers of personal data processing involved in the training, fine-tuning and offering of companion chatbots to scrutinise them through the lens of data protection by design (Section 5.3).

5.1.Data protection by design, fountain of youth

The material scope of Articles 24(1) and 25(1) GDPR–which, as argued in earlier work,[71] should be read together–is broader than what their wording suggests. Article 24(1) requires controllers to "implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with *this Regulation*", while Article 25(1) states that they shall "integrate the necessary safeguards into the processing in order to meet the requirements of *this Regulation* and protect the rights of data subjects" (emphasis added). At first sight, the reference to the requirements of "this Regulation" in both provisions seems to favour a restrictive reading of their material scope of application. Yet, that conclusion only holds true if what the GDPR "requires" is, in fact, limited to complying with the finite set of principles and rules it contains. Article 1(2), which sets the objective of the GDPR, recalls that the Regulation aims to "protect [the] fundamental rights and freedoms of natural persons and *in particular* their right to the protection of personal data". Recital 4 adds that the Regulation "observes the freedoms and principles recognised in the Charter as enshrined in the Treaties, *in particular* the respect for private and family life, home and communications, the protection of personal data, freedom of thought, conscience and religion, freedom of expression and information, freedom to conduct a business, the right to an effective remedy and to a fair trial, and cultural, religious and linguistic diversity". What the GDPR "requires", then, is that controllers mitigate *all* the risks to data subject's fundamental rights *arising from* the processing of their personal data.

This ties back to the nature of the GDPR as a legislative instrument, i.e., a piece of secondary EU law that operationalises that overarching goal by laying down rules to protect *all* natural persons' fundamental rights, *including but not limited to privacy and data protection*, in the context of *the processing of their personal data*. This suggests that "data protection" can either refer to the set of implementing rules contained in Directives and Regulations, or to its fundamental right component. While the recognition of data protection as an independent fundamental right in Article 8 CFREU has led some authors to question

[70] See, for a more detailed analysis of the enforcement architecture of the AI Act: Nathalie A Smuha and Karen Yeung, 'The European Union's AI Act: Beyond Motherhood and Apple Pie?' https://papers.ssrn.com/abstract=4874852.

[71] For an in-depth overview of the history and material scope of data protection by design, I refer the reader to the author's earlier work on the topic, and more specifically to: Pierre Dewitte, 'A Brief History of Data Protection by Design: From Multilateral Security to Article 25(1) GDPR' [2023] Technology and Regulation 80 https://techreg.org/article/view/13807; Pierre Dewitte, 'Fifty Shades of Impact Assessment: An analysis of data protection by design in the case law of national supervisory authorities', [2024] Technology and Regulation (forthcoming).

its exact added value,[72] the EU legislator considered it sufficiently important to warrant a dedicated mention in Article 16 the Treaty on the Functioning of the European Union. Bottom line being, the GDPR, and therefore the "appropriate technical and organisational measures" that controllers must implement pursuant to Articles 24(1) and 25(1), should not only strive to protect data subject's fundamental right to data protection–whatever it adds to the EU fundamental right ecosystem–but, more importantly, also guarantee the respect for *other* fundamental rights such as privacy, freedom of thought, freedom of expression, non-discrimination or cultural, religious and linguistic diversity.

The EDPB has positioned itself in favour of that broad interpretation when it stated that "the data protection principles are in Article 5 (henceforth 'the principles') [and] the data subjects' rights and freedoms are the *fundamental rights and freedoms of natural persons*, and in particular their right to the protection of personal data, whose protection is named in Article 1(2) as the objective of the GDPR (henceforth 'the rights')".[73] And so has the EDPS, when it underlined that "the assets to protect are the individuals whose data are processed and in particular their *fundamental rights and freedoms*" (emphasis added).[74] In that sense, putting the "risk" in "risk-based approach" will *always* require *a* form of risk management process and, as such, will *inevitably* lead to the implementation of measures that have not been *explicitly* foreseen in the GDPR.[75] Controllers are in the driving seat when it comes to the risk identification and mitigation process, supported by guidance in the form of soft law instruments, and inspired by concrete examples emanating from administrative and judicial case law. As such, data protection by design is also a Swiss knife for national supervisory authorities to gradually shape and orient what is expected from controllers in a wide diversity of scenarios. That flexibility is the essence of the risk-based approach, and what makes the combined reading of Articles 24(1) and 25(1) the "keeper of relevance" of the GDPR.

5.2.A proxy to Fundamental Rights Impact Assessments

The discussion as to the material scope of data protection by design sets the scene for the distinction between the many forms of "by design" obligations and corresponding "impact assessments". If "privacy

[72] Among the authors that have contributed to that debate, see: Bart van der Sloot, 'Legal Fundamentalism: Is Data Protection Really a Fundamental Right?' in Ronald Leenes and others (eds), *Data Protection and Privacy: (In)visibilities and Infrastructures*, vol 36 (Springer International Publishing 2017) http://link.springer.com/10.1007/978-3-319-50796-5_1; Orla Lynskey, 'Deconstructing Data Protection: The "added Value" of a Right to Data Protection in the EU Legal Order' (2014) 63 International & Comparative Law Quarterly 569 https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/deconstructing-data-protection-the-addedvalue-of-a-right-to-data-protection-in-the-eu-legal-order/95BD4CCF4670466FD4F6EBAD7DDB4E76; Gloria González Fuster, 'EU Fundamental Rights and Personal Data Protection', *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer, Cham 2014) https://link.springer.com/chapter/10.1007/978-3-319-05023-2_6; Raphaël Gellert and Serge Gutwirth, 'The Legal Construction of Privacy and Data Protection' (2013) 29 Computer Law & Security Review 522 https://linkinghub.elsevier.com/retrieve/pii/S0267364913001325; Gloria González Fuster and Raphaël Gellert, 'The Fundamental Right of Data Protection in the European Union: In Search of an Uncharted Right' (2012) 26 International Review of Law, Computers & Technology 73 http://www.tandfonline.com/doi/abs/10.1080/13600869.2012.646798.

[73] European Data Protection Board, 'Guidelines 4/2019 on Article 25 Data Protection by Design and by Default' para 11 https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf. "Their precise formulation can be found in the EU Charter of Fundamental Rights", it added.

[74] European Data Protection Supervisor, 'Opinion 5/2018 - Preliminary Opinion on Privacy by Design' para 28 https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf. Unfortunately, the accent is once again put on the general principles. See, more specifically, para 30 which states that "these data protection principles, set out in Article 5, can be considered as the goals to achieve".

[75] I therefore share Raphaël Gellert's position when he regrets the dichotomy often found in DPIA methodologies between strict "compliance" issues on the one hand, and "risk management" aspects on the other. The former encompasses the latter, as substantiating the provisions of the Regulation *in concreto* calls for such a risk assessment. See: Raphaël Gellert, 'Understanding the Notion of Risk in the General Data Protection Regulation' (2018) 34 Computer Law & Security Review 279, 283–284 https://www.sciencedirect.com/science/article/pii/S0267364917302698, more specifically point 3.2.

by design" and "data protection by design" have been used interchangeably during the reform process,[76] the above paragraphs have shed light on the importance to clarify the *object* of the assessment, i.e., what fundamental rights does the processing operations at stake impact, and the *purpose* of the countermeasures to be implemented by controllers pursuant to Articles 24(1) and 25(1) GDPR, i.e., how to appropriately mitigate that impact. This suggests the existence of different forms of "risk assessment" that vary in scope and complexity (see Figure 3).[77] The broadest would be a *Fundamental Right Impact Assessment* ("FRIA"), itself the sum of multiple assessments focusing on the impact of the processing of one's personal data on a *specific* fundamental right. This is in line with the conclusions drawn by Karen Yeung and Lee Bygrave in their cross-disciplinary analysis of the Regulation's architecture, in which they argue that "the risk-based approach necessitates that the data controller undertake a contextual 'fundamental rights risk assessment' in order to identify the appropriate level of stringency of the technical and organizational measures that must be adopted to guard against those risks from materializing".[78]

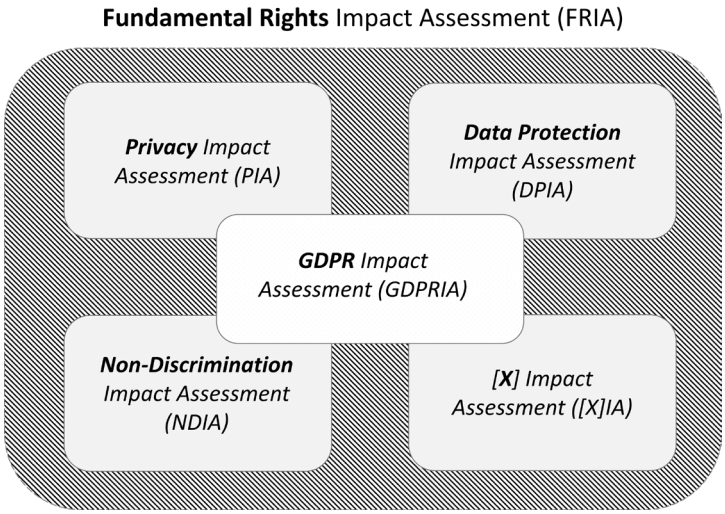

Figure 3: FRIAs, [X]IAs and GDPRIAs

While *privacy* ("PIA") and *data protection* ("DPIA") are the usual suspects, the GDPR strives to protect, as discussed above, *all* data subject's fundamental rights including, for instance, *freedom of expression* ("FoEIA"), *non-discrimination* ("NDIA"), the *right to conduct a business* ("RCBIA") or the *right to an effective remedy a fair trial* ("RERIA"). Or, literally, *any other* fundamental right ("[X]IA"). Building on the role of the GDPR as a "proxy" to mitigate the most pressing risks associated to the processing of personal data for these fundamental rights,[79] performing a *GDPR Impact Assessment* ("GDPRIA")–that is,

[76] With the European Commission, 'Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions - A Comprehensive Approach on Personal Data Protection in the European Union' https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0609:FIN:EN:PDF, for instance, referring to the term "privacy by design" as early as 2010. The European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), COM/2012/011 Final - 2012/0011 (COD)' https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2012%3A0011%3AFIN later settled on the notion of "data protection by design" when drafting Article 23 – which became Article 25 in the final version of the Regulation.

[77] For a longer take on that argument, I invite the reader to consult the author's earlier work: Pierre Dewitte, 'The Many Shades of Impact Assessments – An analysis of data protection by design in the case law of national supervisory authorities' (n 71).

[78] Karen Yeung and Lee A Bygrave, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' (2022) 16 Regulation & Governance 137, 146–147 https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12401.

[79] The EDPS pitched the same idea in European Data Protection Supervisor (n 74) para 61, if with a slightly different meaning, when it stated that "the GDPR looks at [the general principles of Article 5] as goals to achieve, used as 'proxies' to protect individuals' fundamental rights and freedoms, independently of the level of risk".

assessing the degree of compliance of a set of processing operations with the principles and rules it contains, and remedying any deficiency–would lay the groundwork for such a FRIA. While both exercises overlap, the former does not exhaust the latter as controllers will need to complement their compliance efforts depending on the risks inherent to their *specific* activities.

5.3.Personal data, personal data everywhere

Companion chatbots make for a particularly interesting use case. As noted above, the recipe used to cook these products makes use of personal data at nearly every step of the process, from training to fine-tuning to deployment. This, in turn, drags the controllers involved in that algorithmic supply chain into GDPR territory, and requires them to comply with Articles 24(1) and 25(1). As hinted above, these provisions do not *explicitly* prohibit companion chatbot providers from discriminating individuals. Neither do they *oblige* them to prevent their bots from manipulating their users. Rather, they require controllers, as detailed above, to protect *all* data subject's fundamental rights, *including but not limited to privacy and data protection*, in the context of the *processing of their personal data*. In other words, identifying *a* processing of personal data that threatens *any* of the fundamental rights and freedoms enjoyed by data subjects will trigger the obligation for the controllers to mitigate the risk of that threat actually materialising. Overall, and building on the broad interpretation of the material scope of data protection by design put forward in Sections 5.1 and 5.2, I argue that regulators are *already* well-equipped to deal with the issues outlined in Section 3, mainly–if not exclusively–in the form of accountability, data protection by design, and the obligation to conduct DPIAs.[80]

Table 1 provides a high-level, fictional example of the different layers of personal data processing that the development of companion chatbots typically involves. For each layer, the corresponding (joint) controller(s) will have to identify and mitigate the risks these processing operations raise for data subject's fundamental rights, as required by Articles 24(1) and 25(1) GDPR. Layer #0 refers to the personal data originally collected by the entity from which the provider of the training dataset will source its data. Layer #1 covers the scraping of that personal data for the purpose of assembling the training dataset. Layers #2 and #3 capture, respectively, the use of the said training dataset to either train general-purpose LLMs ("GP-LLMs") or fine-tune the behaviour of existing models in specific scenarios ("S-LLMs"). Lastly, layer #4 represents the collection, by the entities responsible for providing the interface through which end-users can access companion chatbots, of the personal data necessary for service purposes.

Having deconstructed the different processing of personal data involved in the development of companion chatbots, the following paragraphs delve into the types of mitigation strategies that controllers could be required, pursuant to Articles 24(1) and 25(1), to implement at each step of the process to address the risks these products create for their users, and society as a whole.[81] Two disclaimers, though. First, the selection of countermeasures discussed below is limited to those that address the specific risks outlined in Section 3, namely biases and discrimination, emotional dependency, and early exposure to sexually explicit content. The very point of this paper is, indeed, to illustrate how data protection by design can serve as a proxy to oblige controllers to go *beyond* the letter of the Regulation, and not to iterate over the many infringements of the GDPR one can criticise providers of companion chatbots for. Second, the selection of

[80] Nathalie Smuha makes a similar point, when she notes that "the AI Act should not be considered as a lex specialis that deviates from data protection rules, but rather as a supplement to fill in the legal gaps that the GDPR, the LED, the EUDPR and other pieces of EU legislation did not yet satisfactorily cover". See: Nathalie Smuha, 'The Paramountcy of Data Protection Law in the Age of AI (Acts)', *Two decades of personal data protection. What's next? EDPS 20th Anniversary* (Publications Office of the European Union 2024) 232 https://lirias.kuleuven.be/retrieve/765234.

[81] It is worth noting that, while the present section starts from the postulate that the risks detailed in Section 3 result from the processing of *personal* data, discrimination, manipulation and harms to mental health may also arise from the processing of *non-personal* data. See, on that: Przemysław Pałka, 'Harmed While Anonymous: Beyond the Personal/Non-Personal Distinction in Data Governance' (2023) 2023 Technology and Regulation 22 https://techreg.org/article/view/13829, specifically Section 3.2.

countermeasures is nowhere near exhaustive. Exactly as the technical overview provided in Section 2, it is rather *instrumental* in making a point as to the "enabling" role of Articles 24(1) and 25(1) GDPR.

Layer	Controller	Processing	Personal data	Purpose
#0	Company A	Collection	User-generated content	Share with people
#1	Company B	Scraping	User-generated content	Create training dataset
#2	Company C	Usage	Training dataset	Train GP-LLMs
#3	Company D	Usage	Training dataset	Fine-tune S-LLMs
#4	Company E	Collection	Account-related data	Provide comp. chatbots

Table 1: Data processing throughout the companion chatbots supply chain

5.3.1. Layer #0. Publicly available, yet not freely reusable

Since the datasets used to train and fine-tune LLMs are usually comprised of publicly available data scraped from the internet, they might contain personal data, especially when the data sources include any form of user-generated content. The "Reddit Conversation Corpus" dataset scraped by Nouha Dziri and her co-authors, for instance, might very-well contain snippets of texts that could lead to the reidentification of the Redditors who originally posted them.[82] The same conclusion holds for "The Pile", the training dataset compiled by EleutherAI that contains,[83] among many other data sources, the "Enron Corpus", which itself holds more than 500.000 emails exchanged by former employees of the company originally made public in the context of an investigation by the Federal Energy Regulatory Commission.[84] The case of Clearview AI is particularly telling. Back in 2020, the company started to scrape the internet, including social media platforms, to gather images and videos to train its facial recognition software and offer its clients–among which law enforcement authorities–a search engine designed to look up individuals on the basis of another picture.[85] OpenAI also scraped vast amount of text data from the internet, including personal data, to train the different iterations of its GPTs, which led the Garante to question the lawfulness of that processing in a series of highly-publicised decisions back in March and April 2023.[86] This is even more problematic considering that adversaries might be able to extract, with

[82] Nouha Dziri and others, 'Augmenting Neural Response Generation with Context-Aware Topical Attention' in Yun-Nung Chen and others (eds), Proceedings of the First Workshop on NLP for Conversational AI (Association for Computational Linguistics 2019) https://aclanthology.org/W19-4103. The raw dataset is available at: https://github.com/nouhadziri/THRED.

[83] A detailed description of the dataset is available in: Gao and others (n 17) figs 1 and 2.

[84] Bryan Klimt and Yiming Yang, 'The Enron Corpus: A New Dataset for Email Classification Research' in Jean-François Boulicaut and others (eds), *Machine Learning: ECML 2004*, vol 3201 (Springer Berlin Heidelberg 2004) http://link.springer.com/10.1007/978-3-540-30115-8_22. As pointed out by Gao and others (n 17) s F.22, the Enron dataset contains information such as former employees' full name, telephone number, as well as the content of their emails in the clear. One of which was sent by a certain Carol St. Clair who wrote "I want to make sure that my vacation time gets paid at 100% before I go down to the 90% level. Thanks for taking care of this. As you can see, I now have access to my e-mail so when I'm not pumping, feeding, changing diapers".

[85] After multiple complaints before different national supervisory authorities and a surge in media attention, Clearview AI was fined by the Italian (Garante per la protezione dei dati personali, *Ordinanza ingiunzione nei confronti di Clearview AI* https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9751362), Greek (Αρχή προστασίας δεδομένων προσωπικού χαρακτήρα, *Επιβολή προστίμου στην εταιρεία Clearview AI* https://www.dpa.gr/el/enimerwtiko/prakseisArxis/epiboli-prostimoy-stin-etaireia-clearview-ai-inc), French (Commission Nationale de l'Informatique et des Libertés, *Délibération SAN-2022-019 du 17 octobre 2022* https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000046444859) and UK regulators (Information Commissioner's Office, *Monetary penalty notice to Clearview AI* https://ico.org.uk/media/action-weve-taken/mpns/4020436/clearview-ai-inc-mpn-20220518.pdf) for having unlawfully processed these images. The Austrian regulator has recently issued a similar decision, if not paired with a fine. See: Datenschutzbehörde, Decision of 9 may 2023 against Clearview AI https://noyb.eu/sites/default/files/2023-05/Clearview\%20Decision\%20Redacted.pdf.

[86] See the Garante's decisions to temporarily ban, then conditionally reauthorize, ChatGPT on the Italian territory: Garante per la protezione dei dati personali, *Provvedimento del 30 marzo 2023 [9870832]* https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832; *ChatGPT: Garante privacy,*

relatively simple queries, part of the dataset used to train these LLMs.[87] Provided that such training dataset contained personal data, this would allow attackers to leverage LLMs' propensity to "memorise" and "regurgitate" random aspects of their training datasets to recover portions of that information.[88] On 29 January 2024, the Italian regulator went even further and notified OpenAI that it had found several breaches of the Regulation,[89] but would take into account the output of the–at the time of writing–ongoing EDPB task force on the matter.[90]

As a side note, it is also unclear whether scraping publicly accessible personal data should be regarded as a *further processing* activity subject to the compatibility assessment pursuant to Articles 6(1)b and 6(4) GDPR, or as a *new collection* for which the said entity would automatically need to rely on a *different* lawful ground than the one used to legitimise the original collection. The recent guidance issued by the Dutch regulator on the matter seems to favour the latter approach, as it states that "[t]he possibility of compatible use is in principle limited to further processing of personal data *by the controller itself* within its own business operations" (emphasis added, free translation).[91] In other words, the implication of *another entity* than the one responsible for the collection would *prevent* that entity from *even invoking* the non-incompatibility of its activities. Yet, severing the link between the purpose specified for the collection and that of the subsequent usage of that data on the sole ground that another entity enters the data processing chain seems, *prima facie*, difficult to reconcile with the rationale of purpose limitation, i.e., inhibiting mission creep and fostering predictability by "prevent[ing] the use of individuals' personal data in a way (or for further purposes) that they might find unexpected, inappropriate or otherwise objectionable".[92] In practice though, the impact of that distinction might be limited, as the elements that would factor in the non-incompatibility test pursuant to Article 6(4) are likely to be integrated within the three-step test of Article 6(1)f–more specifically its "balancing" component–, which is the only lawful ground controllers can reasonably rely on to legitimise the scraping of publicly accessible personal data.[93]

Regardless, one could consider scraping as one of the earliest–if far from the only–causes of the risks raised by companion chatbots at the very end of the supply chain, since it fuels the training and fine-tuning

---

limitazione provvisoria sospesa se OpenAI adotterà le misure richieste L'autorità ha dato tempo allá società fino al 30 aprile per mettersi in regola https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751; *ChatGPT: OpenAI riapre la piattaforma in italia garantendo più trasparenza e più diritti a utenti e non utenti europei* https://www.gpdp.it/home/docweb/-/docweb-display/docweb/9881490.

[87] As recently demonstrated in Milad Nasr and others, 'Scalable Extraction of Training Data from (Production) Language Models' http://arxiv.org/abs/2311.17035. It is worth noting that, according to the authors, "larger and more capable models" such as GPT-4 "are more vulnerable to data extraction attacks".

[88] In Nasr and others (n 87), the authors were, for instance, able to extract phone numbers, email addresses and physical addresses (see Figure 5). They concluded that "16.9% of generations [they] tested contained memorized PII, and 85.8% of generations that contained potential PII were actual PII".

[89] Garante per la protezione dei dati personali, *ChatGPT: Garante privacy, notificato a OpenAI l'atto di contestazione per le violazioni alla normativa privacy* https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020.

[90] The EDPB announced the creation of the task force back in April 2023. See: https://www.edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt. In May 2024, it published a meager interim report documenting the results of the said taskforce that "reflect[s] the common denominator agreed by the Supervisory Authorities in their interpretation of the applicable provisions of the GDPR in relation to the matters that are within the scope of their investigation". See: European Data Protection Board, 'Report of the Work Undertaken by the ChatGPT Taskforce' https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf. Also relevant is the EDPS' first oritentation on the matter. See: European Data Protection Supervisor, 'Generative AI and the EUDPR. First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems' https://www.edps.europa.eu/system/files/2024-05/24-05-29_genai_orientations_en_0.pdf.

[91] Autoriteit Persoonsgegevens, 'Richtlijnen Scraping Door Private Organisaties En Particulieren' 10 https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf.

[92] Article 29 Working Party, 'Opinion 03/2013 on Purpose Limitation' 4, 11 https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[93] I refer the reader to delve deeper into the matter to consult: Catherine Altobelli and others, 'To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR', Deep Driving into Data Protection - 1979-2019 Celebrating 40 Years of Privacy and Data Protection at the CRIDS (2021) https://zenodo.org/record/6411788.

of the LLMs these products all rely on. Translated into data protection terms, this means that the *initial* collection of personal by the controllers responsible for these online resources (layer #0) *already* raises the risk of third party scraping (layer #1), and therefore plays a role, even if limited, in the training and fine-tuning of language models (layers #2 and #3) and the risks they raise for end-users as detailed in Section 3 (layer #4). As such, one could argue that data protection by design would require these controllers to implement technical and organisational measures to *prevent* scraping from *even happening* in the first place, or at least *limit its scope*, where they are aware that third parties tap into their databases for the purpose of assembling training datasets.

That reasoning, however, calls for two comments. First, the causal relationship between the collection of personal data by, say, Facebook, and the risk of manipulation raised by a companion chatbot fine-tuned using a training dataset comprised of posts scraped from that platform, is rather thin. While the processing in layer #0 might *influence* the said risk by making data publicly accessible, it is far from the only–or even main–factor that determines the likelihood of that risk materialising, or its severity for the affected individuals. As such, one cannot reasonably require controllers to anticipate *all* the potential risks that their processing operations could *contribute* to raise, especially where their role in shaping the said risks is limited. The determining factor to gauge the extent of controllers' responsibilities under Article 24(1) and 25(1) GDPR, I argue, is a combination of *awareness of* and *agency over* the risks at stake. The following paragraphs delve deeper into those aspects.

Second, it is worth noting that, besides the risks identified in Section 3, scraping raises other concerns for individuals, including the risk of targeted cyberattacks, identify fraud, hyper-personalised political targeting, spamming and overall loss of control over one's personal data. Now the role that social media platforms play in the materialisation of these risks, compared to those associated to the use of companion chatbots trained or fine-tuned using personal data scraped from these platforms, is more prevalent since these risks are the *direct* consequences of the public availability of that data. The limitations inherent to the reasoning deployed above are therefore less relevant, and one could more convincingly leverage data protection by design to oblige social media platforms to implement technical and organisational measures to protect personal data from unlawful scraping. A joint statement from the International Enforcement Cooperation Working Group argues in that direction, and recommends "social media companies" to implement "multi-layered technical and procedural controls" such as "rate limiting", "monitoring how quickly and aggressively new accounts starts looking for other users", "taking step to detect scrapers by identifying patterns in 'bot' activity" and "using CAPTCHAs".[94] Illustrating the overlap between GDPRIAs and FRIAs introduced in Section 5.2, one can argue that the obligation for social media platforms to go to such lengths follows from *both* the specific provisions on security and data breaches (Articles 5(1)f, 32-34 GDPR) *and* the broader obligation to identify and mitigate all the risks raised by the processing of personal data for data subject's fundamental rights and freedoms, including but not limited to privacy and data protection.

5.3.2. Layer #1. Assemble the finest training datasets

As a direct result of the above, the datasets used to train or fine-tune the LLMs that companion chatbots all rely on might also contain personal data. *If* that is the case and *if* the assembler purposefully crafts the said dataset for others to train their LLMs on, one could argue that Articles 24(1) and 25(1) GDPR would require that entity to proactively identify and mitigate the issues that could lead the resulting models to, for instance, discriminate against certain communities, or behave in a manipulative way. The "appropriate technical and organisational measures" that the assembler could be required to implement would include, *a minima*, the provision of a detailed description of the content, structure and biases of the training

[94] International Enforcement Cooperation Working Group, 'Joint Statement on Data Scraping and the Protection of Privacy' https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf.

dataset.[95] One might argue that sanitisation techniques could be used to filter the personal data out of the training dataset. These are, however, no silver bullet since (i) the sanitisation process itself would amount to a "processing" of "personal data", therefore triggering the applicability of the GDPR for that limited portion of the processing, and (ii) sanitisation techniques work best for identifying well-formatted information, such as social security numbers, while determining what constitutes "personal data" within the meaning of Article 4(1) GDPR is a context-sensitive rather than formalistic exercise.[96] Measures could also include debiasing techniques to remedy glaring flaws in the dataset, though some of the risks outlined in Section 3 can also be attributed to the broader social, political and economic context in which the resulting models are deployed.[97] One could even leverage data protection by design to argue that the assembler should refrain from creating datasets the *sole purpose* of which is to train language models designed to discriminate or manipulate people, or violate their fundamental rights and freedoms in any other way. That, however, might prove particularly tricky to demonstrate.

Of course, that obligation comes on top of all the other requirements stemming from the Regulation such as, when it comes to the scraping of publicly accessible personal data, the obligation to rely on one of the six lawful grounds listed in Article 6(1) GDPR.[98] On that point, it is worth noting that the CNIL recently held a somewhat debatable position regarding the lawfulness of the reuse of training datasets assembled by third parties. In its fourth "AI how-to sheet", the French regulator indeed states that "[c]ertain failures committed by the controller to set up and disseminate a dataset do not *systematically* and *irreparably* affect the lawfulness of the processing carried out by the re-user". "Thus", it adds, "a re-user may use a dataset whose illegalities are *minor*, provided that the reuse meets the requirements of the GDPR". As such, argues the CNIL, the entities reusing existing training datasets can limit their assessment to checking whether "there is *no clear doubt* that the dataset is lawful (in particular that the source processing is not manifestly lacking of a legal basis when the data are so intrusive that they cannot be processed without the consent of the individuals), ensuring in particular that the conditions for collecting the data are sufficiently documented" (emphasis added).[99] What the CNIL understands by "minor illegality" and "no clear doubt" is, unfortunately, not detailed in the guidelines.[100] This, in my opinion, is a rather slippery slope as it paves

[95] It is worth noting that Article 10(2)f and g of the AI Act mandates that the training, validation and testing data sets used in the context of high-risk AI systems be checked for and cleared from "possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law".

[96] Hannah Brown and others, 'What Does It Mean for a Language Model to Preserve Privacy?', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) 2287 https://dl.acm.org/doi/10.1145/3531146.3534642.

[97] In that sense, debiasing only addresses a subset of the issues raised by AI systems, and by extension companion chatbots. On that, see: Seda Gürses and Agathe Balayn, 'Beyond Debiasing - Regulating AI and Its Inequalities' (European Digital Rights (EDRi) 2021) Report https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/.

[98] It is worth highlighting that the Garante has already relied on a breach of the lawfulness principle for the training of the LLMs used in both Replika and ChatGPT to ban these services on the Italian territory. See, respectively, Garante per la protezione dei dati personali, *Provvedimento del 2 febbraio 2023* https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214; *Provvedimento del 30 marzo 2023* https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832. The Italian regulator later reauthorised ChatGPT to resume service in the country after OpenAI implemented additional controls. See (n 86). For an overview of these features, see the dedicated Help Centre Article on OpenAI's website: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. While OpenAI now offers users the possibility to opt-out from their conversations being used to fine tune GPT-4, the company has yet to come up with any solution to remedy the unlawfulness of the processing of the personal data contained in the dataset used to train its LLM. OpenAI now faces a class action in California for a breach of both data protection and copyright law. See: Gerrit De Vynck, 'ChatGPT Maker OpenAI Faces a Lawsuit over How It Used People's Data' *Washington Post* (28 June 2023) https://www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/.

[99] See Commission Nationale de l'Informatique et des Libertés, 'Ensuring the lawfulness of the data processing - In case of re-use of data, carrying out the necessary additional tests and verifications' https://www.cnil.fr/fr/node/164402.

[100] I still ought to refer the reader to the guidance offered in Commission Nationale de l'Informatique et des Libertés, 'Projet de Guide Pratique - Ouverture et Réutilisation de Données Publiquement Accessibles' https://www.cnil.fr/sites/cnil/files/2023-08/projet_de_guide_ouverture_partage_et_reutilisation_de_donnees.pdf, especially the "Fiche principe n°2 : Comment identifier la base légale de son traitement ?" and the "Fiche cas d'usage n°3 : la réutilisation de données publiquement accessibles à des fins de constitution ou d'enrichissement de fichiers destinés à la prospection commerciale". The guide is, unfortunately, only available in French, and will likely be reviewed following a public consultation opened on 1 August 2023.

the way for controllers to invoke vague excuses to not pay too much attention to the provenance and lawfulness of the datasets they select and use to train or refine their own models.

Yet, sanitisation and debiasing suffer from several limitations. The first is related to the objective of data protection by design, which is to mitigate the risks raised by the processing at stake for *data subject's* fundamental rights and freedoms. Meaning, in the context of the scraping of personal data to assemble training datasets, the risks for the individuals *whose personal data are included in the said datasets*. These risks might be very different from the risks raised by the processing of personal data of *other individuals* by the *other actors* involved in the companion chatbots supply chain that might have used that dataset to train their own models, or fine-tuned a pre-trained model that has itself relied on that dataset. This is particularly problematic in cases where the *absence* of a certain community from a given training dataset is the *very source* of the discrimination risk. The broader the scraping, the more "data subjects" are likely to be affected, though. Maybe up to a point where the sheer amount and diversity of the individuals included in the training dataset warrant to move past the notion of "data subject" and instead consider the risks to society as a whole. That interpretation finds support in the wording of Articles 24(1) and 25(1) GDPR, if one pays closer attention to the exact vocabulary used by the EU legislator. Karen Yeung and Lee Bygrave also seem to vouch for that interpretation.[101] Indeed, while the measures that controllers must implement should strive to protect "the rights of *data subjects*", the "risks" that controllers must "take into account" when selecting which of these measures are appropriate in a given scenario are those for "the rights and freedoms of *natural persons*". The latter is, without contest, broader than the former.

The second has to do with the limited influence of the assembler on the potentially harmful processing happening further down the chain. The mere fact that a training dataset *can* be used to train a language model that will end up discriminating or manipulating its users does not trigger the obligation for its assembler to anticipate and prevent *all* the scenarios in which these risks might materialise themselves in practice. That reasoning would only hold water if, as hinted at above, the *sole* purpose of the training dataset *is* to train harmful LLMs. PygmalionAI's PIPPA dataset, which contains more than 1 million lines of dialogue between users of Character.ai and its large language model and is mainly used to train LLMs able to spit out explicit content,[102] is a case in point. Can that training dataset be used by companies such as Chai Research Corp. to fine-tune over-sexualised chatbots made accessible to underage users without any age verification mechanism?[103] Sure. Does that mean that PygmalionAI has to mitigate that specific risk? Probably not, as chatbot-powered ERP is not, *in itself*, problematic. How it is implemented by providers of LLMs, downstream developers and, most importantly, providers of companion chatbots, and what safeguards these actors put in place to frame its usage, are the critical questions.

5.3.3. Layer #2. Cook general-purpose LLMs

Going one step down the companion chatbots supply chain, the exact same reasoning, *including* its limitations, can be held against the entities that use these datasets to train their own LLM. Building on the argumentation detailed above, *if* the training dataset contains personal data and *if* the provider of the LLM determines the purposes and the means of the training process, then Articles 24(1) and 25(1) GDPR will require the implementation of "appropriate technical and organisational measures" to mitigate the risks it

[101] The authors note that "controllers who must carry out these assessments may be ill-equipped to do so for at least four partially overlapping reasons", the second of which is the fact that they "are required to consider fundamental rights of *individuals* generally, rather than merely the rights of data subjects directly implicated by the proposed processing" (emphasis added). Yeung and Bygrave (n 78) 147.

[102] On HuggingFace, PygmalionAI warns that "PIPPA contains conversations, themes and scenarios which can be considered 'not safe for work' (NSFW) and/or heavily disturbing in nature" and that "models trained purely with PIPPA may have the tendency to generate X-rated output". See the dataset card available at https://huggingface.co/datasets/PygmalionAI/PIPPA.

[103] This is one of the issues that pushed the Garante to temporarily ban ChatGPT on the Italian territory. The risk of early exposure to sexually explicit content was also at the core of the Garante's earlier decision to ban Replika on the Italian territory, as bots on the platform were serving "utterly inappropriate replies" to children "having regard to their degree of development and self-conscience". See, on these two decisions, (n 86).

poses for *data subject's* rights and freedoms. Taking into account, here again, the broader risks for "the rights and freedoms of *natural persons*". Although this time one might argue that, since the entity that trains the language model actually knows what it will be used for–i.e., conversing with people in the case of an off-the-shelf model, or serving as baseline for the fine-tuning of a model that will eventually be used to do so–, these measures should account for the *specific* risks raised by the use of such conversational agents. With greater awareness of the purpose for which the model will be used, comes a better understanding of the risks it is likely to pose for end-users. The more conscious the controller, the heavier the burden of implementing "appropriate" countermeasures. These could include, for instance, proper testing to ensure that the model does not output toxic speech, racist slur or manipulative patterns inherited from the training dataset(s), and the implementation of fixes to limit that behaviour as much as possible.

Exactly as for the providers of training datasets, however, it would be unreasonable to ask from providers of LLMs that they anticipate *all* the scenarios in which their models might endanger data subject's fundamental rights. But the interpretation of data protection by design defended in Sections 5.1 and 5.2 would certainly bar providers of LLMs from releasing *intrinsically* harmful models without any form of mitigation for the risks their usage might pose for end-users. Data protection law set aside, this also follows from a "moral" obligation not to unleash damaging products in the wild. Illustrating the above, the developers of "DialoGPT", a tuneable neural conversational response generation model trained by researchers at Microsoft, have acknowledged that their model "retains the potential to generate output that may trigger offense", to "reflect gender and other historical biases", to "express agreement with propositions that are unethical, biased or offensive", or to "disagree with otherwise ethical statements".[104] But they spontaneously tried to clean the training dataset from overtly offensive data before moving on with the training process, without referencing any of the potentially applicable regulatory frameworks.

The risks raised by these models encouraged "marketplaces" such as GitHub, HuggingFace and Civitai to progressively roll-out content policies designed to ban users from uploading certain types of models, either based on their *intended*, *actual* or *potential* harmful uses. Yet, as rightfully noted by Robert Gorwa and Michael Veale in a recent case study focusing on these three platforms, the peculiar nature of models as "content containing other content", and the resulting uncertainties regarding the applicability of the EU intermediary liability framework, makes moderating models a thorny policy challenge.[105] These content policies come on top of the various flavours of licences concluded between model uploaders and users, which also often limit the use that the latter can make of these models.[106] In both cases, however, enforcement of these contractual clauses remain the exclusive prerogative of platforms and licensors.

5.3.4. Layer #3. Spice things up with a hint of fine-tuning

Moving even closer to end-users, downstream developers can also fine-tune existing pre-trained language models to behave in a certain way. For instance, next to the PIPPA dataset, PygmalionAI also offers Pygmalion 6B,[107] a proof-of-concept dialogue model based on EleutherAI's GPT-J 6B primarily designed to act as a NSFW role-play partner. On the dedicated HuggingFace page, PygmalionAI states that the dataset used to fine-tune GPT-J 6B "consisted of 56MB of dialogue data gathered from multiple sources, which includes both real and partially machine-generated conversations". Another such example includes the "uncensored" version of LLaMA 2 13B, specifically retrained with a filtered dataset to "reduce

[104] Yizhe Zhang and others, 'DIALOGPT: Large-Scale Generative Pre-Training for Conversational Response Generation' in Asli Celikyilmaz and Tsung-Hsien Wen (eds), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics 2024) https://aclanthology.org/2020.acl-demos.30.
[105] Robert Gorwa and Michael Veale, 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries' [2024] Law, Innovation and Technology (forthcoming) 13–17 https://osf.io/6dfk3.
[106] See, for instance, the OpenRAIL family of licences that seeks to "prevent irresponsible and harmful applications" of algorithms, code and data. See, for sample licenses: https://www.licenses.ai/ai-licenses.
[107] As noted in Section 2, Pygmalion 7B and 13B are now available.

refusals, avoidance, and bias".[108] Here again: *if* that dataset contains personal data, and *if* PygmalionAI qualifies as the controller for the fine-tuning process, it will have to comply with Articles 24(1) and 25(1) GDPR and assess the risks that the processing poses for *data subject's* fundamental rights and freedoms, and potentially for all *natural persons* if one sticks to the broader interpretation of these provisions. In this case, PygmalionAI unequivocally warns that their model "is *not* suitable for use by minors" and "*will* output X-rated content under certain circumstances" (emphasis in original).

As detailed above, the fact that the company has *specifically* developed Pygmalion 6B for ERP and *explicitly* acknowledges the risks that this might pose if the model is used to engage with minors drastically raises the bar, in my opinion, with regard to the countermeasures that PygmalionAI should implement to prevent these risks from materialising. These should at least include mechanisms to ask users their age, and appropriately tone down or even deactivate the generation of NSFW language. Or to detect certain prompts that might reveal that the individual interacting with the fine-tuned model is, in fact, underage.[109] There is only so much downstream developers can do, that is. Self-declarations of one's age or date of birth has never proven a particularly efficient technique to prevent children from accessing sexually explicit content online, and the same observation can probably be made when it comes to chatbot-prompted equivalents. Besides, Pygmalion 6B, like many other fine-tuned models, is not gated behind an API but is available for anyone to download under a GNU General Public License.[110] As a result, PygmalionAI does not retain any form control over the way chatbot developers implement, fine-tune and market the model in practice. Knives don't kill people, right?

5.3.5. Layer #4. Serve half-baked in a shiny user interface

This brings us to the most critical link of the chain: the entities that *actually* offer companion chatbots to end-users. As hinted at in Section 2, these actors can intervene throughout the entire supply chain, or simply propose the technical infrastructure necessary to engage with existing language models. The determining factor is the provision of the *interface* through which individuals can access, personalise and converse with companion chatbots, regardless of the degree of influence that the said entity had over the different building blocks outlined in Figure 2. That interface reflects design and marketing choices that influence the type of audience that the companion chatbot will eventually attract.

The processing of personal data at stake here is radically different from that involved in the scenarios outlined above, as these entities collect and further process the personal data of *end-users* for the purpose of *providing companion chatbot services*, as opposed to that of the natural persons that might have been included in the datasets used for training or fine-tuning purposes. In the case of the Chai application, for instance, that includes account-related data, usage data, conversation histories and bot parameters, the qualification of which as "personal data" within the meaning of Article 4(1) GDPR makes little doubt. *If* the entity that develops and offers the user-facing interface acts as the controller for the above-mentioned processing–which I argue it does, either as a sole or joint controller–, it will also have to comply with the many requirements stemming from the Regulation. But this time, the risks for data subject's' fundamental rights are drastically different–and so are their likelihood and severity.

In our complaint against Chai Research Corp., the background of which is detailed in Section 6, we flagged glaring infringements of the GDPR including a breach of the lawfulness and transparency principles, and the lack of any mechanism to verify that children's consent is either given or authorised by

[108] The model is accessible here: https://huggingface.co/ehartford/WizardLM-1.0-Uncensored-Llama2-13b. Note that the developers clearly warn that "An uncensored model has no guardrails" and that "You are responsible for anything you do with the model, just as you are responsible for anything you do with any dangerous object such as a knife, gun, lighter, or car".
[109] Reference is made, for instance, to the advice that Snapchat's My AI gave someone impersonating a 13 years old girl on how to make her first time with a 31 years old man special: by "setting the mood with candles or music". See: Fowler (n 62).
[110] See, for the exact terms of the license: https://www.gnu.org/licenses/gpl-3.0.en.html. See, for the original Twitter post by Tristan Harris, cofounder of the Center For Humane Technology: https://twitter.com/tristanharris/status/1634299911872348160/photo/3.

the holder of parental responsibility as required by Article 8(2). But, more importantly, we noted the absence of any "technical and organisational measure" to ensure that the processing does not infringe data subject's fundamental rights and freedoms.[111] This is all the more troublesome given that (i) the use of companion chatbots raises specific issues for their users, (ii) these risks are abundantly documented in field-oriented and academic literature, (iii) data subjects are the end-users, and end-users are the individuals at risk, which evacuates the question as to whether the objective of Articles 24(1) and 25(1) GDPR includes the protection of "natural persons" more generally, (iv) the target audience of companion chatbots is often comprised of vulnerable individuals, which increases the likelihood and severity of these risks and, (v) the proximity with end-users puts the provider of the user-facing interface in the driver's seat when it comes to the implementation of appropriate countermeasures.

In light of the above, I argue that a broad interpretation of Articles 24(1) and 25(1) GDPR, as defended in Sections 5.1 and 5.2, requires developers of companion chatbots to identify and mitigate *all* the issues raised by the services they put on the market, as long as these result from the processing of their users' personal data. Vertically-integrated developers could address the risk of discrimination by, for instance, debiasing their training dataset,[112] moderating the output that can universally be regarded as discriminatory,[113] or contextualising debatable statements. They could also detect early signs of emotional dependency through text classification techniques, at which point they could redirect the user to relevant resources or even human support.[114] When it comes to preventing premature exposure to sexually explicit content, they could pair mandatory age verification mechanisms with filtering techniques to shield younger users from ERP. In a surprisingly bold business blog post, the FTC recently called upon chatbots providers to stop "misrepresenting what these services are or can do" and refrain from placing them on the market without "adequately mitigating risks of harmful output".[115] The list goes on, and providing a comprehensive overview of these measures extends well beyond the scope of the present paper.

The main problem is, the vast majority of companion chatbots developers has implemented none–or too little–of that. Likely because doing so would run contrary to either their business model, which often relies on premium subscription plans or advertising revenues, or to the purpose for which some of these chatbots are mainly used. One can't help but question whether the objective of Romantic AI, for instance, is really to serve as a sparring partner to "train personal communication skills in romantic and love areas" or is nothing more than a paid-for NSFW text generator appealing primarily to the male fantasy.[116] Or to discern a subtle hint of irony when the developers of Tavern AI, a platform notoriously used for ERP, answer the question "Can this technology be used for sexooo?" by "Surprisingly, our development team has received reports that some users are indeed engaging with our product in this manner. We are as

[111] On that note, it is worth noting that Chai's "AI Safety Framework" looks more like a public relationship move than an actual attempt at making its fine-tuned model "safe" for users, as the company measures "safety" *solely* in terms of the average amount of "Note Safe For Work" (typo in original) words the model outputs per day. Which is strange at best, since Lit-6B, the model used by Chai Research Corp. to fine-tune GPT-J 6B, is *specifically* designed to generate NSFW language. See, on that framework: Xiaoding Lu and others, 'The Chai Platform's AI Safety Framework' 4 http://arxiv.org/abs/2306.02979.

[112] Recalling, once again, that debiasing is no silver bullet. See Gürses and Balayn (n 97).

[113] OpenAI has, for instance, developed a moderation tool that allow downstream developers to check whether the text generated by its LLMs complies with its usage policies. More specifically, the tool is able to detect hate speech, harassment, self-harm, sexual and violent content. See: https://platform.openai.com/docs/guides/moderation/overview. The same could be developed by providers of companion chatbots to address the risks discussed in Section 3.

[114] Michael Tadesse et al. have, for instance, demonstrated the ability of word embedding to detect suicidal thoughts. See: Michael Mesfin Tadesse and others, 'Detection of Suicide Ideation in Social Media Forums Using Deep Learning' (2020) 13 Algorithms 7 https://www.mdpi.com/1999-4893/13/1/7. UNICEF has also developed a list of common trigger words used by young users when disclosing abuse or risks of harm. See: UNICEF East Asia Pacific, Gender section., 'Safer Chatbots Implementation Guide' (UNICEF 2023) Report https://www.unicef.org/documents/safer-chatbots-implementation-guide.

[115] Michael Atleson, 'Succor Borne Every Minute' (*Federal Trade Commission Business Blog*, 7 June 2024) https://www.ftc.gov/business-guidance/blog/2024/06/succor-borne-every-minute.

[116] Not to mention the risk that a chatbot built around the motto "Wanna be macho? She will be stunning!" is likely to raise for "natural persons" and society as a whole by shaping unrealistic and sexist expectations of what one should expect from a "romantic" encounter.

puzzled by this as you are, and will be monitoring the situation in order to gain actionable insights".[117] A hint of addiction might go a long way in keeping these services afloat, one might daresay. Again, the point of this section is *not* to condemn ERP as such, or companion chatbots altogether. Rather, it underlines the needs for developers to provide a healthy and safe environment for their deployment and use, taking into account the documented risks these tools might raise for end-users. Where companion chatbots involve the processing of personal data, data protection by design is a powerful proxy to force controllers to, at the very least, initiate a reflection as to the safeguards that should be put in place to mitigate these risks.

6. The Safe AI Companion Collective ("SAICC") initiative

The risks outlined in Section 3 are but a handful of the issues raised by companion chatbots. Many more are documented in practice, but are not discussed in the present paper for obvious length reasons.[118] This led me to co-author an open letter with Nathalie Smuha, Mieke De Ketelaere, Mark Coeckelbergh and Yves Poullet in March 2023 urging regulators and policymakers to set up awareness campaigns to shed light on the issues surrounding companion chatbots, and encouraging developers to proactively identify and mitigate the risks they pose for individuals.[119] Education, we argued, has a crucial role to play in alleviating some of these concerns, as does a wider public debate on the function we wish companion chatbots–as well as AI systems more generally–to serve in our society.

The interdisciplinary work undertaken in the context of that open letter laid the foundation for a more structured collaboration in the form of the Safe AI Companion Collective ("SAICC"), a platform that I co-created with Nathalie Smuha (Legal scholar and philosopher at the KU Leuven Faculty of Law), Mieke De Ketelaere (Associate Professor on Sustainable, Ethical and Trustworthy AI at the Vlerick Business School) and Thomas Ghys (Privacy expert and CEO of Webclew) to (i) raise awareness about the risks of shaping intimate relationships through insufficiently tested and controlled AI systems, (ii) advocate for the proper enforcement of data protection and consumer protection legislation against providers of training datasets, LLMs and companion chatbots, and (iii) share selected resources to help the public navigate the technical, societal, legal and ethical implications of these services. The platform, which is accessible at the address https://www.saicc.info/, was launched in summer 2023 and disseminated to stakeholders ranging from policymakers to academics, journalists and representatives from the industry.

[117] See TavernAI's FAQ: https://github.com/TavernAI/TavernAI/blob/main/faq.md. The same document also redirects users to websites on which they can download pre-made characters, even though it acknowledges that "these sites are filled to the brim with weird shit. Like, you'll be lucky if half the characters aren't furry, or even alive".

[118] For a comprehensive overview of the risks raised by LLMs and companion chatbots, I refer the reader to the following resources: Weidinger and others (n 27); Forbrukerrådet (n 26); Grant Fergusson and others, 'Generating Harms - Generative AI's Impact & Paths Forward' (Electronic Privacy Information Center 2023) Report https://epic.org/wp-content/uploads/2023/05/EPIC-Generative-AI-White-Paper-May2023.pdf; Daniel Leufer and Méabh Maguire, 'What You Need to Know about Generative AI and Human Rights' (*Access Now*, 24 May 2023) https://www.accessnow.org/what-you-need-to-know-about-generative-ai-and-human-rights/.

[119] The Open Letter has been published in English, French, and Dutch. See: Nathalie Smuha and others, 'Open Letter: We Are Not Ready for Manipulative AI – Urgent Need for Action' (*KU Leuven AI Summer School Blog*, 29 March 2023) https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai; Nathalie Smuha and others, 'We Are Not Ready for Manipulative AI – Urgent Need for Action' [2023] *Euractiv* https://www.euractiv.com/section/digital/opinion/we-are-not-ready-for-manipulative-ai-urgent-need-for-action/; Nathalie Smuha and others, 'Le chatbot Eliza a brisé une vie : il est temps d'agir face à l'IA manipulatrice' *La Libre* (29 March 2023) https://www.lalibre.be/debats/2023/03/29/le-chatbot-eliza-a-brise-une-vie-il-est-temps-dagir-face-a-lia-manipulatrice-BSGGRV7IBRDNROO33EWGFVMWAA/; Nathalie Smuha and others, 'Onze samenleving is niet klaar voor manipulatieve AI' [2023] *Knack* https://www.knack.be/nieuws/technologie/onze-samenleving-is-niet-klaar-voor-manipulatieve-ai/.

Within the context of SAICC, we also filed a complaint against Chai Research Corp., the company behind the Chai platform, before the Belgian supervisory authority for several infringements of the GDPR.[120] While we could have targeted other providers of companion chatbots, we felt that Chai Research Corp.'s community-based development model, which effectively shifts part of the burden to ensure AI safety and compliance onto independent developers who are rewarded for optimising user engagement, raised particularly salient issues. A machine-translated version of the complaint originally submitted in French, as well as all the evidence filed before the authority, is included as supplementary material to the present paper and is accessible here: https://doi.org/10.48804/K6RDSG. The complaint details the infringements attributable to Chai Research Corp., including breaches of the lawfulness and transparency principles, of the rules surrounding children's consent, and of the obligation to conduct a DPIA where the processing concerns "vulnerable data subjects" and the "innovative use or application of new technological or organisational solutions" possibly "with a high risk to individuals' rights and freedoms".[121]

More specifically, we built our argumentation around the absence of *any* process designed to identify and mitigate the risks raised by Chai's activities for its users' fundamental rights and freedoms, including but not limited to safety, privacy and data protection. Such exercise, the outcome of which should be communicated to data subject in part or in full, is critical to ensure that all users enjoy a safe and healthy experience when using companion chatbots. While the primary goal of our initiative is to force Chai Research Corp. to implement the necessary guardrails to mitigate the impact of their chatbots, the complaint also serves an academic agenda. That is, to demonstrate that data protection by design can serve as a powerful proxy to force controllers to proactively identify and mitigate some of the risks that are not *explicitly* addressed in the Regulation, but are nonetheless the result of *a* processing of personal data. That reasoning hinges on the broad interpretation of the material scope of Articles 24(1) and 25(1) GDPR put forward in Sections 5.1 and 5.2, and the phase-oriented approach introduced in Section 5.3, to address the risks outlined in Section 3.

This, we hope, will help put the issue of companion chatbots on regulators' agenda, and pave the way for a decision to condition the availability of these tools to a prior and comprehensive assessment of the many risks they pose for their user base. We also wish to raise awareness among individuals, policymakers, regulators and developers on the need to consider the impact of technology on people's live before rushing to the market. Not only is this "by design" approach becoming an integral part of the European legislator's response to the challenges raised by emerging technologies, but it is also instrumental in avoiding that individuals become the first victims of half-baked products and services. Lastly, we believe that our methodology might inspire other people to launch similar initiatives in other countries. This might help elevate the issue at the EU level, potentially through the possibility offered to data protection authorities to "request that any matter of general application or producing effects in more than one Member State be examined by the Board with a view to obtaining an opinion" (Article 64(2) GDPR). This is crucial, as the risks associated to the use of personalised chatbots such as those offered by Chai Research Corp. extend well beyond Belgium, and considering that similar services continuously flood the market.

7. Conclusions

Wrapping up, the broad interpretation of data protection by design put forward in Sections 5.1 and 5.2 paves the way for leveraging the risk-based approach to oblige the different controllers that *directly* or *indirectly* contribute to shaping the likelihood and severity of a given risk to identify and proportionally

[120] The Litigation Chamber of the Belgian Autorité de Protection des Données informed us on 14 June 2024 that it had decided to ask its Inspection Service to open a formal investigation onto the matter. It also noted that, after querying other supervisory authority through the cooperation mechanism, no other regulator had received a similar complaint.

[121] See, respectively, points 7 and 8 of Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679' 10 https://ec.europa.eu/newsroom/document.cfm?doc_id=47711.

mitigate that risk. Yet, doing so in the context of companion chatbots is particularly challenging since the supply chain of their development and deployment (i) is comprised of multiple layers of processing operations that all contribute, to some extent, to raising specific risks, and (ii) involves a wide range of actors, each of which is responsible for a subset of these processing activities. This makes Articles 24(1) and 25(1) GDPR compelling but challenging options to try and address the risks outlined in Section 3. Condensing the findings presented throughout this paper, the following sections highlight the challenges of applying data protection by design in complex supply chains (Section 7.1), and the role of awareness in broadening the scope of controllers' risk management exercise (Section 7.2).

7.1.Data protection by design in complex supply chains

For data protection by design to unlock its full potential, two cumulative conditions must be met. First, and since Articles 24(1) and 25(1) GDPR only apply to "controllers", it requires a precise understanding of the role of each actor involved in the companion chatbots supply chain. This is essential to identify who "determines the purposes and means" of the processing at stake, and allocate responsibilities for the implementation of "appropriate technical and organisational measures". As detailed in Section 2.2, companion chatbots are rarely the product of a single entity. Translated into data protection terms, this means that there might be multiple (joint) controllers involved at various stages of the production, deployment and use of companion chatbots for different, yet interdependent sets of personal data processing operations. That concerns extends to AI systems in general.[122]

In the words of Jennifer Cobbe and her co-authors, it is indeed "no longer necessarily true that computer systems are produced by a group of developers or an organisation, or by a vendor simply integrating various standalone components into one product". Instead, computer systems "now often involve a group of organisations arranged together in a data-driven supply chain, *each retaining control over component systems they provide as services to others*" (emphasis in original).[123] Companion chatbots are prime examples of such data-driven supply chain products. Against that background, bringing the issue of data protection by design to the table *also* sparks an essential reflection as to the broader ecosystem in which each individual actor operates, which is the first step toward the implementation of *appropriate* countermeasures, as discussed below. Doing so might partially alleviate the issue documented by David Gray Wider and Dawn Nafus in their interviews with 27 AI engineers across different industries, i.e., that few of them *felt* like the risks such as the ones discussed in Section 3 fell within their agency, capability, or responsibility to address.[124] Besides, "controllership" within the meaning of Article 4(7) GDPR does not *always* fall upon the entity that *factually controls* the technology at stake.[125]

Second, and for the "technical and organisational measures" to fulfil their goal of protecting "data subject's fundamental rights and freedoms", all the controllers involved in the AI supply chain should *ideally* be aware of the purpose and context in which the final product–in this case, companion chatbots–will be deployed. That, of course, gets more difficult the more actors are involved in the chain. As hinted

[122] This is often referred to as the "many hands" problem, already pitched by Helen Nissenbaum back in the nineties. See: Helen Nissenbaum, 'Accountability in a Computerized Society' (1996) 2 Science and Engineering Ethics 25 https://doi.org/10.1007/BF02639315.

[123] Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Understanding Accountability in Algorithmic Supply Chains', Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery 2023) https://doi.org/10.1145/3593013.3594073. Another characteristic of algorithmic supply chain, they add, is that "certain key actors–in particular, major cloud providers who often control underlying technologies–provide many services to millions of customers, holding important positions across supply chains in many sectors".

[124] David Gray Widder and Dawn Nafus, 'Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility' (2023) 10 Big Data & Society 20539517231177620 https://doi.org/10.1177/20539517231177620.

[125] Cobbe, Veale and Singh (n 123) 7–8, also highlight that disconnect. While they state that "those who are factually responsible for various aspects of production, distribution, and use of algorithmic systems must be identified correctly so that accountability can be allocated accordingly", they also acknowledge that the "assignment of legal roles and responsibilities does not describe the real interdependencies and power relations between AI service providers […] and their customers".

at above, assemblers of training datasets, for instance, are simply not in a position to anticipate all the scenarios in which their datasets might be used, and the corresponding risks that these training activities might pose for data subjects. Similarly, developers or general-purpose LLMs such as the ones shared on HuggingFace might not always be aware that downstream entities use them as basis to fine-tune language models to generate sexually explicit content, let alone be factually able to address the issues raised by this type of chatbots. Jennifer Cobbe et al. refer to the "accountability horizon", i.e., "the point beyond which an actor cannot 'see', which depends on the actor and the chain", and acknowledge the problem that is poses for legislative initiatives that rely on "impact assessment, risk assessment and risk management mechanisms to mitigate the harms of AI technologies" such as the GDPR. The proper identification and mitigation of these risks, further note the authors, therefore "require knowledge of both the AI technology's specification and development and the purpose and context of its application".[126]

7.2.Broadening the "accountability horizon"

Once aware, each actor must then implement technical and organisational measures to mitigate the risks–or the portions thereof–that can be attributed to the processing of personal data for which they qualify as controller. As detailed in Section 5.3, multiple controllers, through their own successive or concurrent processing activities, can contribute to shaping the characteristics of a specific risk. This could be the case, for instance, for the risk that an LLM trained and fine-tuned by different entities outputs biased or discriminating content when implemented within a user-facing companion chatbot. In that case, all the actors that have been involved in the training of the underlying model would, I argue, be compelled to implement appropriate countermeasures *in proportion to their respective contribution to the risk*. On the contrary, controllers the processing operations of which have *not influenced that risk* would not be required to include it within the scope of their risk management exercise. One could think, for instance, about the risk of early exposure to sexually explicit content, for which the provider of a general-purpose LLM, even if fine-tuned by downstream developers, assumes no or limited responsibilities. Along the same line, the burden of mitigating the risk of emotional dependency associated to the use of companion chatbots would primarily fall on the providers of these tools, as it is mainly, if not exclusively, their part of the processing that raises that risk in the first place.

This calls for two remarks. First, the "accountability horizon"–and therefore the extent of the risk management exercise–broadens proportionally to the degree of awareness of a given actor as to the purpose for and context in which its contribution will or is likely to be used. As a result, if an entity becomes aware of a risk that its individual input propagates up or down the supply chain–such as a general-purpose language model "A" showing extreme biases towards certain demographies when deployed in service "B"–, that actor will, if one agrees with the conclusion drawn in Section 5, have to implement appropriate measures to mitigate that risk. In the context of companion chatbots, the closer the actor is to end-users, the more tangible the risks, and the more extensive the countermeasures it must implement pursuant to Articles 24(1) and 25(1) GDPR. This is the rationale behind an idea pitched earlier in this paper, i.e., that the entity that fine-tunes an existing LLM *specifically* for ERP, for instance, faces a heavier burden in terms of implementation of mitigation measures than an entity that further trains it to partially automate customer service tasks.

Second, and provided that the entities involved in the companion chatbot supply chain qualify as "controllers" for at least *a* "processing" of "personal data", the broad interpretation of the material scope of data protection by design defended in Sections 5.1 and 5.2 requires them to look *beyond* their own

[126] Cobbe, Veale and Singh (n 123) 9. "Yet", they add, "without advance knowledge of their customers' many, varied, and changing application contexts and uses, providers cannot properly account for the range of potential risks that might arise. Similarly, without knowledge of or influence over production, customers cannot reliably assess how systems are developed, nor ensure that systems are appropriate to the risks arising in their context. Even where they have some knowledge, models are regularly updated, and customers may lack visibility or capacity to reassess. In many cases, therefore, no actor will have sufficient knowledge of or control over both production and deployment to be able to reliably assess or mitigate the impacts and risks".

processing activities and also consider the risks that *downstream* or *upstream* usage of their respective contribution might pose for individuals. In that sense, I argue, data protection by design is a solid *argument* to compel controllers to "expand their accountability horizon", to use the wording of Jennifer Cobbe et al., if not an actual *solution* to foster a better understanding of the interdependencies between the many actors involved in the AI supply chain. That reading essentially turns Articles 24(1) and 25(1) GDPR into a binding obligation for controllers to take the broader ecosystem in which their processing operate into account, and to break away from the tunnel vision typically associated to independent actors. Data protection by design is, in that sense, inextricably linked to the allocation of responsibilities.

I would even go as far as defending that Articles 24(1) and 25(1) GDPR, by requiring each actor to consider the impact of its individual contribution to the overall supply chain, foster a deeper, more fundamental conversation as to the *desirability* of systems such as companion chatbots. In that sense, data protection by design, precisely because it calls upon controllers to assess the risks that their processing operations might pose for *data subject's* fundamental rights, also taking their impact on *natural persons* into account, acts as a first line "sanity check" to assess the broader societal, political and economic implications of AI systems.[127] It is now up to national supervisory authorities and courts to exploit the full potential of data protection by design as one–if not the only–tool at their disposal to address the risks raised by companion chatbots, and AI systems more generally.

[127] Such is the question raised by the "second wave of algorithmic accountability" research, as coined by Frank Pasquale, 'The Second Wave of Algorithmic Accountability' (*LPE Project*, 25 November 2019) https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/, that focuses on "whether [AI systems] should be used at all–and, if so, who gets to govern them". See, on that point: Julia Powles and Helen Nissenbaum, 'The Seductive Diversion of "Solving" Bias in Artificial Intelligence' (*OneZero*, 7 December 2018) https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.