ETHICAL IMPLICATIONS OF AI-POWERED CHATBOTS: CONSIDERATIONS FOR THE AI ACT: A CASE STUDY OF TESSA

Dawn Branley-Bell / Johannes Feiner / Sabine Prossnegg

Dawn Branley-Bell, Associate Professor, Northumbria University, Department of Psychology, Northumbria University,
Newcastle upon Tyne, NE1 8ST, UK, dawn.branley-bell@northumbria.ac.uk

Sabine Prossnegg, Associate Professor (FH), FH JOANNEUM, Institute for Software Design and Security
Werk-VI-Strasse 46, 8605 Kapfenberg, AT; Sabine.Prossnegg@fh-joanneum.at; https://www.fh-joanneum.at

Johannes Feiner, Senior Lecturer (FH), FH JOANNEUM, Institute for Software Design and Security
Werk-VI-Strasse 46, 8605 Kapfenberg, AT; Johannes.Feiner@fh-joanneum.at; https://www.fh-joanneum.at

Keywords: ***AI, chatbots, eating disorders, responsibility, ethics, risk assessment***

Abstract: *The use of Artificial Intelligence (AI), including the use of chatbots, is common and prevalence is expected to continue to rise. This paper delves into the creation and deployment of a chatbot named Tessa. Tessa was intended to aid users' self-assessment of symptoms indicative of eating disorders and guide them towards relevant support services. The chatbot was designed to help ease strain on overburdened healthcare staff and offer support for individuals who may face significant delays in being able to access an in-person medical consultation. Unfortunately, despite a promising start, a recent incident with Tessa demonstrated how chatbots can go wrong. This paper analyses the incident from technical, psychological, and legal viewpoints, with a specific focus on key considerations around responsibility and safeguarding of chatbots within the health domain and the AI Act. This paper contributes to the ongoing discourse on the implications of AI-driven healthcare interventions, fostering a critical dialogue for future developments in this evolving landscape. We support the idea of regular assessments of AI interventions, improved regulation, and more stringent consideration of ethical and safeguarding issues.*

1. Introduction

In recent months, a chatbot named Tessa has made many headlines[1]. Tessa was a chatbot designed with the intention to support individuals vulnerable to eating disorders (EDs). After it's development, the chatbot was adopted by the National Eating Disorders Association (NEDA) where it was officially launched and introduced as a 'wellness chatbot' in February 2022. The researchers behind the original development of Tessa stressed that they appreciate the need for human interaction, and they did not design the chatbot to be a replacement for human services such as NEDA's telephone helpline. However, controversy followed when NEDA subsequently shut down its helpline in June 2022 after 20 years of operation.[2] Unfortunately, further controversy hit when the chatbot went on to provide users with harmful dieting advice. Tessa was subsequently discontinued from use in May 2023.[3]

[1] "An eating disorders chatbot offered dieting advice, raising fears about AI in health", Updated June 9, 2023, by Kate Wells; NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice by Staff Writer June 5, 2023.

[2] https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/, last accessed 14.6.2023, https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea, last accessed 14.6.2023.

[3] https://www.cbsnews.com/news/eating-disorder-helpline-chatbot-disabled/, last accessed 14.6.2023.

It is important to recognise that no technology is inherently good or bad. AI has the potential for huge benefits across many areas of our lives, healthcare is one of them. One such form of AI application in healthcare is the use of chatbots. A chatbot is an AI-based computer program designed to mimic human-to-human conversation by analysing the user's text-based input, and providing smart, related answers [Dahiya, 2017]. Chatbots are increasingly being used within healthcare [Softić et al. 2021], where they use the information provided by the user to extract information about their health concerns and the best course of action for treatment or further investigation. Chatbots have the potential to help reduce burden on notoriously overworked healthcare staff, reduce patient wait times, increase accessibility and scope of healthcare services and support, and reduce costs. Chatbots can potentially provide an encouraging 'first step' for individuals who may feel negative emotions (e.g., embarrassment, perceived stigmatisation) which prevent them from accessing traditional healthcare services [Branley-Bell, 2023]. A recent systematic review of research into potential healthcare applications of ChatGPT (and similar technologies) suggests that chatbots have the potential to revolutionise healthcare delivery [Muftić et al. 2023]. Similar results have been found in a recent scoping review looking at the application of chatbots for anxiety and depression [Ahmed et al. 2023]. EDs have the highest mortality rate of all mental health conditions but current healthcare wait times are extreme[4] and individuals experiencing EDs face significant, potentially devastating, delays for support. Chatbots have significant potential benefits in this area. Most individuals with EDs never seek support and feelings of embarrassment or shame have been cited as some of the reasons why this may be the case.

However, as with any technology – particularly one that is in its relative infancy – there are limitations and challenges to overcome. This is particularly important when technology is applied to sensitive topics and with consequences relating to health and wellbeing. Some of the main challenges include the acceptability of AI within society; explainability of AI and how the system works; concerns around accuracy and bias; lack of clarity around where responsibility lies; and concerns around potential for human replacement.

2. Technical Perspective – What may have gone wrong?

As the source code of Tessa is not freely available, we can only speculate about the inner working. According to reports, Tessa was designed to only have a limited number of pre-determined responses and both the researchers and NEDA were keen to emphasise that the chatbot was not based on ChatGPT or Generative Learning Models (GLM), and that it was not able to 'off piste' or generate responses that deviated from its pre-programmed replies [Chan et al. 2022]. This is often referred to as a rule-based system, whereby system answers are calculated in a predictive way by following predefined rules set up by domain experts. Although this has also been criticised as impersonal, others have likened it to the strict scripts that human volunteers are provided with when manning telephone helplines. The major problems occurred when somehow Tessa switched from a rule-based chatbot approach to a generative language approach. How this occurred is unclear. To foster understanding we try to explain the technical limitations and implications of software systems built on artificial intelligence before discussing its use within Tessa.

AI has been around for a significant period, but it wasn‘t always in the spotlight until ChatGPT came along. Nevertheless, we‘ve become accustomed to its applications in our daily lives, such as recommender systems when shopping online, text auto-suggestions as we type, and smartphones that enhance images using computational optimisation. Machine learning (ML) constitutes just one facet of the vast field of artificial intelligence. ML itself is further divided into various subcategories, with supervised, unsupervised, and reinforcement learning being among the most well-known. Sometimes, these different types are used in conjunction, creating hybrid systems. Supervised learning involves utilising labelled data to make predictions using a trained model. Unsupervised learning, on the other hand, helps identify patterns or clusters within unlabelled

[4] https://www.nuffieldtrust.org.uk/resource/children-and-young-people-with-an-eating-disorder-waiting-times.

data. In the trial-and-error style of reinforcement learning, an agent receives feedback or rewards from its environment to enhance its learning process.

To explain, *supervised learning models* are like teaching a computer with labelled examples. Imagine you are teaching a child to recognise different animals. You show the child pictures of cats, dogs, and birds, and you tell them what each animal is. Over time, the child learns to recognise these animals by their features. Similarly, in supervised ML, we give a computer lots of examples (like pictures of animals) and tell it what they are (labelling them). The computer learns to make predictions based on these labelled examples. So, when you show it a new picture, it can tell you if it‘s a cat, dog, or bird because it learned from the labelled data. In comparison, *unsupervised learning* is like asking the computer to solve a jigsaw puzzle without a picture on the box. Imagine you have a jigsaw puzzle with pieces, but you don‘t know what the final picture should look like. In unsupervised learning, the computer tries to group similar puzzle pieces together based on their shapes and colours. It doesn‘t know what the whole picture is, but it sees that some pieces fit well with others. This helps it to discover patterns and relationships in the data. So, unsupervised learning helps us organise and make sense of data without having pre-labelled examples. Lastly, *reinforcement learning* is a bit like training a dog. Imagine you have a pet dog, and you want it to perform tricks. When your dog does a trick correctly, you give it a treat as a reward. If your dog does something wrong, it doesn‘t get a treat. Over time, your dog learns which actions lead to rewards (treats) and which actions don‘t. In reinforcement learning, we give the computer program a task, and it tries different actions to figure out which one lead to more rewards and which ones lead to fewer rewards, therefore learning to make better decisions over time.

GLMs are designed to generate human-like text or speech responses. The GLM is trained on vast amounts of data (e.g. articles, websites, books) during which it learns patterns and relationships about this data. The final resulting model is capable of selecting and/or assembling answers based on predictions, i.e., the system using the model decides which answer to give based on the current (and previous) input. That is good and bad at the same time. On the one hand ML-systems are, after the initial development effort, easy and cost-effective to run. They always generate different answers, so interaction is more like a human conversation. However, it is also important to recognise their limitations:

AI systems are not always correct, although they may often appear objective to users. Assumptions can be made that they will always give scientifically grounded answers, but this is not the case.

AI systems can also change over time. Many AI models may include systems that are designed to enable them to improve through use. Retraining and fine-tuning as part of an iterative process can be beneficial (and indeed recommended to keep information up to date). However, it also provides an opportunity for the system to deteriorate or ‘misbehave’ if there are errors or biases in the new training data.

AI systems often lack explainability and transparency: Explainability is necessary to help individuals using AI systems to critique the output they are presented with. Without a chain of arguments, or references to sources, it is not easy to evaluate an output and know how much trust to put into the given answers. For example, a system might report skin cancer for a provided image. Then it is crucial to give reasons for the decisions. A lack of explainability can lead to under or over trust in the system. Transparency is also a key requirement and should start with forcing a system to tell users immediately that they are interacting with a computer system and not with human beings.

AI systems do not promote reproducibility. This is particularly true of GLM models like ChatGPT. Many of these models are purposefully built to be more human like, and therefore unique responses are often desired by the developers. Even for the exact same input, a different output is generated. Whilst this may be desirable during the chatbots intended use, it can cause problems. For example, user trust in a system can be eroded if differing answers are provided for the same input, this is mainly the case if the user perceives responses to be conflicting in nature. Furthermore, if a user claims to have received an incorrect/unethical/damaging response from the system, as it is often impossible to identify the exact answer that was provided due to the lack of reproducibility.

AI systems are vulnerable to malicious attacks such as hacking. It is possible for someone to intentionally feed (and thereby retrain) a system in a bad way with fake and/or manipulated training data. That might lead to wrong and even dangerous answers. Repairing an attacked system is not easy – sometimes even impossible. You cannot easily tell the system to *forget* learned behaviours.

AI systems may raise questions over privacy as often user input is recorded and processed to improve the data set and models. When sensitive data is uploaded, the authorship, privacy rights and copyright might not be clear. Unfortunately, and despite of the clear obligations within the GDPR, privacy by design, differential privacy and anonymisation are currently seldom considered when systems are built. This also leads to ethical issues around users' rights.

AI systems enforce the tendency towards monopolies when only few large companies collect data about and from users worldwide. Their advantage of possessing large data sets increases over time. Competitors without access to the large data collections are losing out quickly.

3. Impact on human trust in AI

Human trust in technical systems, often referred to as technology trust, is a multifaceted and crucial aspect of our interaction with various technological tools, devices, and systems [Branley-Bell et al., 2020; Lukyanenko et al., 2022]. Trust plays a significant role in determining how individuals, organisations, and societies adopt and use technology. Trust is a psychological and emotional state of positive expectations that a person, organisation, or system will perform in a certain way. In the context of technology, it is the belief that a technical system will function as intended and not lead to negative consequences.

There are many factors that influence trust including perceived reliability (the system‘s ability to consistently perform its intended function without failure or error), security (user beliefs that their data and privacy are protected from unauthorised access), transparency (the information provided to users to aid their understanding of how a system works, its algorithms, and its decision-making processes), usability (how easy a system is to use and understand) and past experience with similar technologies. Incidents such as the issues with Tessa have the potential to erode trust and rebuilding trust can be a challenge [Lukyanenko et al., 2022]. This is particularly true given that this incident occurred when the chatbot was being deployed by an organisation that would often be perceived as trustworthy and reputable – therefore if something like the Tessa incident can occur even in this situation, the impact on public trust in similar technologies is likely to be impacted even further. This could be problematic for encouraging adoption of future AI-based interventions, even if these interventions are suitably regulated. To mitigate against any further impacts on public trust in responsible technologies, it is important to ensure that any future applications are strictly scrutinised for adequate safeguarding.

It would also be beneficial to educate users on the technology they are interacting with, including its capabilities, limitations, and potential risks. This can not only help build trust but can also protect again potential over-trust and ensure responsible use [Buçinca et al., 2021]. So far we have discussed factors which can reduce user trust. However, humans can also exhibit a bias towards *over trusting* technology, particularly if they have not had any previous negative experiences with the technology. This is referred to as automation bias, i.e., the tendency to favour decisions made by automated systems over those made by humans. However, as has been demonstrated, AI systems are also prone to error and users must be encouraged to be more critical of their outputs [Buçinca et al., 2021].

In summary, human trust in technical systems is a complex and dynamic relationship. Building and maintaining trust in technology is essential for its widespread adoption and acceptance. Users‘ trust can be cultivated through a combination of reliability, security, transparency, usability, and ethical considerations, as well as through legal and regulatory mechanisms that promote responsible technology development and use.

4. Legal Perspective – Tessa and the AI Act

Now we turn to consideration of Tessa in the light of the AI Act. We do so with some limitations, like knowing that the AI Act is currently just a draft, knowing not all details about Tessa and picking out just some important topics of the AI Act to deal with[5].

4.1. The AI Act

The EU wants to promote the uptake of human-centric and trustworthy AI systems while at the same time guaranteeing a high level of protection. Like with previous legislative acts of the EU relating to digitisation,[6] it relays heavily on a sound risk management system to achieve these two rather contradictive targets at the same time. The protection of natural persons as outlined in the AI Act comprises health, safety, fundamental rights, democracy and the rule of law. Companies should be encouraged to invest in innovation and find better conditions for the development and use of such technologies.[7]

The AI Act defines AI as a machine-based system that is designed to operate with varying levels of autonomy and that can generate outputs (Art 3 para 1). To distinguish AI from other software, some degree of independence of actions from human controls and of capabilities to operate without human intervention is required.[8] The EU has opted for a horizontal regulation for AI systems, based on proportionality and a risk-based approach complemented by the Code of Conduct for AI systems that pose a high level of risk. That implies that a thorough risk assessment has to be carried out before the application of any AI system, because the higher the risk to individuals' fundamental rights or safety, the greater the system's obligations [Klaushofer 2022].[9]

General principles for the use of AI systems are laid down in Art 4a. The general principles are human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental well-being.

Currently, the AI Act provides for four risk levels. Top level are unacceptable risks like Social Scoring or language assisted toys for children (see Art 5 of Parliaments Version). Those, rather rare applications, are prohibited not only within the EU, but also outside the EU. The prohibited practices relate to the entire market, from placing such AI tools on the market, to commissioning or use.[10] The second risk level comprises high risk application (Art 2) and refers to a critical use in a critical sector. Some uses are critical in all sectors (e.g., recruitment processes) other possible applications include critical infrastructure, safety components (e.g., surgery), profiling for (self)-employment, migration and asylum, democracy and judicial system. High risk AI requires operators to complete a conformity assessment, registration in a database before use and a declaration of conformity. At the moment with the AI Act still in an early stage, the list of high risk uses can and will be

[5] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_de, 9.8.2023, see Art 1 para 1 AI Act, Amendment 3 ff Proposal for a regulation Recital 1 ff AI; all following citation are taken of the P9_TA(2023)0236 Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

[6] https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence, last accessed 9.8.2023.

[7] The Commission issued three legal initiatives regarding trustworthy AI https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence, last accessed 9.8.2023.

[8] See also Amendment 18 Proposal for a regulation Recital 6.

[9] Art 3 para 1 point a and b AI Act.

[10] Recital 10 EU Parliament June 2023, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, last accessed 10.10.2023.

changed later. Health care is explicitly listed as a high-risk use category (recital 37)[11] but it remains unclear if this just affects the access to health care services as such.

Other AI applications like Generative AI or chatbots are assumed to bear only limited risks. The most important requirements listed for this risk level are (1) disclosure that AI is used, (2) a model to prevent generation of illegal content by the system and (3) a summary concerning the use of copyright data.

Finally, there are AI systems that bring only a minimal risk with them. This should be the majority of AI systems, the EU names on its website the use of free of charge applications like AI supported video games or spam filters. These range outside the scope of the AI Act since they only bring minimal or no risks for citizens' rights or for security.[12] Open-source AI systems are also listed as excluded from the AI Act, unless intended for high risk uses (Art 2 para 5e).

4.2. Chatbot Tessa within the AI Act

The first question is whether Tessa is considered as an AI system or a mere software. Chatbots are specifically listed as AI, so we jump to the next question where to put Tessa in the risk pyramid, Art 5 ff AI Act.

Since Chatbot Tessa gives advice to people with EDs, we have several questions to clarify. We can exclude the lowest and highest risk levels, because it certainly is not one of the unacceptable uses and it also can't be a minimal application with no obligation at all. Chatbots are particularly mentioned as limited risk. Described as a 'wellness chatbot' – and not a medical device – it seems that Tessa may not quite meet the criteria for the high-risk category. However, if Tessa indeed falls within the limited risk level, this raises concerns as the chatbot was designed to help a vulnerable population seek help about a serious health condition[13]. The AI Act states that the use of AI systems that decide about the access to and enjoyment of certain essential private and public services, including healthcare services, deserve special consideration. However, the Tessa chatbot may be regarded as "just" an advice tool, and not an essential health service. Having a limited set of answers just like people on a telephone with a predefined list of answers before them, using a conventional risk / severe risk matrix[14] and given that the EU's AI pyramid assumes that most applications are located on the lower end, we assume this would place Tessa, at least based on the rule-based system on which it was originally designed, in the lower limited risk level.[15] We assume that the introduction of the chatbot would be regarded as appropriate due to the immense pressure on health budget and staff. Supplementing the existing human services with the chatbot service could allow easier access for users. Considering its changes, the question arises if the AI Act could have prevented the negative incident that occurred with Tessa giving harmful diet advice.[16] This is with the caveats that it is not possible to know the exact technical details about Tessa and knowing that the AI Act is not in force yet and may still be amended.

The **transparency obligation** (Art 52) was met as users were aware that they were using a 'wellness chatbot' and not talking with a human being. That enables them to take an informed decision to continue or step back.[17] The labelling obligation could be linked to the fact that content is wholly or partly created by AI, as well as the way in which the content is created [Höch 2023]. That said, we think that this transparency obligation has to include substantial changes as well. In our case we rate the change from a fixed set of predefined answers to

[11] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_de, last accessed 11.10.2023; (Pollirer 2023).

[12] See previous footnote.

[13] Recital 67, Annex III para 1 point an and ba, "with regard to the eligibility to heath care, point c: classification or health care triage systems."

[14] Outlined ie in Art 3 para 1 point 1a and 1b AI Act.

[15] Recital 32; Excellence and trust in AI, EU Commission, February 2020.

[16] Even high-risk AI applications require usually both, a critical sector and a critical use. The risk assessment is carried out partly as suggested by [Pollirer 2023].

[17] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 9.8.2023; [Höch 2023].

an open generative system as substantial. Even developers cannot be sure of the outcome. So that transparency obligation could had been breached in our case at least in that respect, that changes have been carried out and how the new algorithm works.

That brings us to the next issues of explainability, transparency and reproducibility of the system itself, namely, is the functioning of the algorithm understandable, a description of input and output data, the logic and the impact of the calculation and can answers be reproduced. However, these things are in many cases not even possible. To the best of our knowledge, we are not aware of this information being provided to the users of Tessa. Unfortunately, this means that some of the core requirements and care obligations of the AI Act have been disregarded with Tessa. On the one hand, the researchers with their special know-how about the functioning, the possibilities and the possible risks would have a special obligation to inform the organisation not only about the existing system but also about the dangers of introducing changes or updates. We must be clear that we do not know if the researchers did provide this information to the organisation who deployed the chatbot – although they do state the following in their 2022 paper:

“First, a regular review of chatbot transcripts is necessary to identify bugs and inappropriate conversations. We believe regular review is necessary even when the program has been finalized, as it is possible that technical issues or issues impacting user experience could be introduced unexpectedly after a change is made. This recommendation is consistent with Beaudry et al, who noted that significant time and costs are incurred in developing and maintaining mental health chatbots.” [Chan 2023].

It must also be recognised that once an intervention has ‘left the hands’ of the researchers, further monitoring is usually not possible because research organisations are trapped in a project centred business model, whereby work on the project stops when funding ends. Initial papers from the researchers indicate that, at least in the beginning, the chatbot appeared to be helping individuals [Fitzsimmons-Craft EE et al. 2022]. Published papers suggest that the chatbot was ‘rule based’, although it is interesting to note that in their 2022 paper the researchers do refer to their chatbot as a ‘rule-based chatbot with features of AI’ and do allude to the possibility of more deep learning processes in future chatbots [Chan et al. 2022].

Our conclusions are that the main responsibility appears to lay with the organization deploying the chatbot. The organization was involved from the development stage so presumably they knew at least about some of the possible risks and didn’t act prudently when introducing the subsequent changes. However, NEDA subsequently blamed the issues on the company that operated Tessa as a free service and that may have changed Tessa without NEDA’s awareness or approval.[18] That company is quoted as saying that changes to Tessa were part of a “systems upgrade” which included generative AI as part of the contract with NEDA. This further demonstrates the complexity of introducing AI systems where multiple stakeholders are involved. The lesson learned from Tessa will have to be that contracts are very important as well as organisational/procedural measures like constant AI system monitoring, assessment and evaluation, even with limited risk chatbots.

5. Recommendations for the AI Act

Transparency is vital to ensure users are aware when they are interacting with computer systems and not a human being, preliminary to achieve trust in AI systems. This includes how data is analysed and/or stored, even more so if personal data are involved. This should be an obligation in the AI Act for all systems (including any further substantial changes to these systems). *Explainability* is needed to allow users to be critical of the AI responses and decisions. A system should present an understandable way for users to see how the given decision was reached. This AI literacy is provided for in the AI Act but it will be quite a challenge to achieve that. *Reproducibility* remains a problem and we do not foresee an easy solution here since even the program-

[18] https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea.

mers cannot predict the answer. However, built-in flagging systems could provide one method to help users record inaccurate and/or potentially distressing/damaging responses. Human oversight, monitoring and also processes for evaluation are important here, as partly envisaged in the AI Act. AI systems are also exposed to *adversarial attacks*. This means, someone can intentionally feed (and thereby retrain) a system with manipulated training data. That might lead to wrong and even dangerous answers. Cybersecurity and safety are big issues these days, there are many rules in place like the GDPR and NIS 2.0[19]. We are a bit critical, because many obligations are already in place and penalties are severe, however, too many regulations confuse companies rather than help them in setting up sound IT systems.

6. Conclusions

AI can bring many benefits and be successfully applied across a wide range of contexts, healthcare is no exception. As reported by the researchers, Tessa was designed with good intention – to help individuals experiencing eating disorder symptoms, and to ease burden on overworked healthcare staff. The goal is admirable but unfortunately things took a turn for the worse on this occasion. However, we must not fear the technology but instead learn how to design, apply, update, monitor and evaluate it in an ethical and appropriate manner. With adherence to the principles set out in this paper, individuals and organisations can strive to ensure they maximise the benefits of adopting AI whilst minimising and mitigating against the risks. Any technology has its limitations, and it is up to us to recognise that and ensure adequate safeguards are in place. As we have seen that legal remedies have to be combined with technical, organisational and ethical measures. We suggest that future versions of AI chatbots (1) undergo repeated intensive technology assessments including extensive testing in many contexts before and during use, (2) provide ways and means to correct decisions (the output reported to users) and the technical underlying systems (ML models engaged), (3) significantly improve explainability to enable users to critically evaluate the system and the responses it provides, and maybe even (4) provide third party inspection of the algorithms and anonymised training data.[20]

Under and over trust of AI systems also poses an issue, again explainability can go some way towards helping to encourage users to more accurately calibrate their trust in a system. Chatbots such as Tessa do have potential benefits, providing that they are deployed and maintained in a responsible manner. Lessons can be learned to mitigate against similar risks in future and to encourage the development of appropriate regulation mechanisms. Finally, AI systems must be tailored to the particular domain or task, in Tessa's case, to ensure that the model behaves in a manner that is appropriate to individuals experiencing symptoms of EDs.

Looking at the current draft AI Act, we see that the obligations could have potentially helped to go some way to prevent the negative consequences associated with chatbot Tessa. We feel that a lot of support will have to be given to companies and users alike to better understand and use new technologies, and also to take a more balanced approach as to whether or not a new technology should be implemented and replace existing systems in the first place.

The lead author's time on this research was funded by the Medical Research Foundation
[MRF-058-0016-F-BRAN-C0868].

[19] A-SIT, BKA, Österreichisches Informationssicherheitshandbuch, Cloud Computing, 15.11.2022.

[20] Dignum, Virginia, Future-proofing AI: regulation for innovation, human rights and societal progress, last accessed 15.6.2023, https://progressivepost.eu/future-proofing-ai-regulation-for-innovation-human-rights-and-societal-progress/, last accessed 11.10.2023; Yaraghi, Niam, ChatGPT and health care: implications for interoperability and fairness, 9.6.2023, https://www.brookings.edu/articles/chatgpt-and-health-care-implications-for-interoperability-and-fairness/ last accessed 11.10.2023.

7. References

Ahmed, Arfan/ Hassan, Asmaa/Aziz, Sarah/Abd-Alrazaq, Alaa/Ali, Nashva/Alzubaidi, Mahmood/Al-Thani, Dena/Elhusein, Bushra/Ali Siddig, Mohamed/Ahmed, Maram/Househ, Mowafa, Chatbot features for anxiety and depression: A scoping review. In: Health Informatics Journal. 29(1) doi:10.1177/14604582221146719 (2023).

Branley-Bell, Dawn/Brown, Richard/ Coventry, Lynne/Sillence, Elizabeth Sillence. Chatbots for embarrassing and stigmatizing conditions: Could chatbots encourage users to seek medical advice. Frontiers in communication. doi: 10.3389/fcomm.2023.1275127 (2023).

Branley-Bell, Dawn/Whitworth, Rebecca/Coventry, Lynne (2020). User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. In: Kurosu, M. (eds) Human-Computer Interaction. Human Values and Quality of Life. HCII 2020. Lecture Notes in Computer Science, vol 12183. Springer, Cham. https://doi.org/10.1007/978-3-030-49065-2_27.

Buçinca, Zana/Malaya, Maja B./Gajos, Krzysztof Z. (2021). 'To Trust or to Think'. Proceedings of the ACM on Human-Computer Interaction. https://dl.acm.org/doi/10.1145/3449287.

Chan, William W/Fitzsimmons-Craft, Ellen E./Smith, Arielle C./Firebaugh, Marie-Laure/ Fowler, Lauren A./ DePietro, Bianca/Topooco, Naira/Wilfley, Denise E./Taylor C. Barr/Jacobson, Nicholas C., The Challenges in Designing a Prevention Chatbot for Eating Disorders: Observational Study. JMIR Form Res 6(1): e28003. doi: 10.2196/28003 (2022).

Dahiya, Menal. A tool of conversation: Chatbot. Int. J. Comput. Sci. Eng. 5, 158–161 (2017).

Fitzsimmons-Craft, Ellen E/Chan, William W/ Smith, Arielle C/Firebaugh, Marie-Laure/Fowler, Lauren A./ Topooco, Naira/DePietro, Bianca/Wilfley, Denise E./Taylor, C. Barr/Jacobson, Nicholas C., Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. Int J Eat Disord. Mar;55(3):343-353. doi: 10.1002/eat.23662. Epub 2021 Dec 28. PMID: 35274362 (2022).

Höch, Dominik/Kahl, Jonas, Anforderungen an eine Kennzeichnungspflicht für KI-Inhalte, KuR 2023, 396–401, 397 (2023).

Klaushofer, Reinhard. „Menschenrechte und KI-Analyse der verbotenen Praktiken im Entwurf zu einem EU-Gesetz über KI" In: Jahrbuch Digitalisierung und Recht, 309-323, 310 (2022).

Lukyanenko, Roman./Maass, Wolfgang./Storey, Veda. C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. Electron Markets 32, 1993–2020 (2022). https://doi.org/10.1007/s12525-022-00605-4.

Muftić Fatima/Kadunić Merjem/Mušinbegović Almina/Almisreb Ali Abd. Exploring medical breakthroughs: a systematic review of ChatGPT applications in healthcare. In: Southeast Eur. J. Soft Computi. 12, 13–41 (2023).

Pollirer, Hans-Jürgen, Checkliste KI und Datenschutz. In Dako, 4, 86–90 (2023).

Softić, Amela/Husić, Jasmina Barakovic/Softić, Aida/Baraković, Sabina. Health chatbot: design, implementation, acceptance and usage motivation. In: 2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH), Bosnia (2021).