Conceptual Framework for the humanization of A.I. called FEA

Dennis Klug
Schwarz IT KG
Economics Department, DHBW Mannheim
Mannheim, Germany
DennisKlug1@googlemail.com

Jan M. Budinger
Boheringer Ingelheim Pharma GmbH & Co. KG
Economics Department, DHBW Mannheim
Mannheim, Germany
jan.budinger@web.de

***Abstract*—With increasingly complex and abstract information systems, more and more often based on A.I., an approach to make these systems easier to access for humans has to be made. The goal should be to increase acceptance and everyday usage of A.I. independent of the user's background to enable our society to exploit all possible application options. To confront this issue, we propose a model called FEA, which makes use of continuous engineering, modifying the development process of A.I. systems and enabling interaction with the user in an iterative shell principle circle. It consists of three layers with the functionality at its core enveloped by the emotional and aesthetical aspects. These adjust the technology to human interaction, with certain requirements needing to be met, until the process can move onto the next layer. The results of a humanizing approach, using for example vocal feedback and assistance in everyday situations, results in a significant increase of acceptance, in all kind of age groups.**

Keywords—Artificial Intelligence, A.I., Computer Science, Humanizing of A.I., FEA, Framework

I. Introduction of FEA

The lack of knowledge about the underlying technology, safety requirements and standards which need to be met by currently active and in development A.I. technologies leads to distrust concerning these technologies. Humanizing of A.I. is a critical element to build trust in this and other increasingly difficult to understand technologies. While most technologies work on the technical side, they have a hard time to appeal to different groups of age, ability and others. To increase the overall acceptance and daily integration of A.I., we propose a model to humanize it. The underlying idea is to make A.I. more natural to interact with, establishing a better and more individual connection with its user, resulting in a higher acceptance independent of the user´s background.

We call this model FEA (Function, Emotion, Aesthetics). With functionality at its core being checked to fulfill certain ethical and legal guidelines, tasks can be accomplished by an A.I. that previously required human involvement, by mimicking human intellect and reasoning in the required task. With the emotional aspect, instead of simply conveying the requested information or the result, the A.I. should be designed to understand and access human speech including figures of speech and expressions, draw conclusions and act accordingly. This increases acceptance and builds trust.

Finally presenting the emotionalized information in an individual as well as aesthetical and appealing manner gives the user the feeling of being understood at a deeper level, matching the user's interests and preferences which helps to build a personal relationship. Also, aesthetically adjustments for different target groups can be made easier, than changing the technology at its fundamental core.

The delimitation of the spectrum covered in this paper, will not focus on how the technical aspects of the model work. Since every algorithm in itself is built to fulfill a different task, as well as different requirements, we focus on the aspect of how we can build an ethical foundation for all algorithms, which are supposed to interact with humans. All examples provided inside this paper, even though supporting the claims inside it, have been conducted without the model we are proposing. Therefore, they are merely present to support our proposed model, instead of showing real applications of it.

II. FEA Model

As seen in Fig. 1, the FEA model follows a shell principle. The technical functionalities and processes form the core enveloped by processes to attach emotions to the technical outputs as well as to present them in an aesthetical and appealing manner to its users. By clearly differentiating between these three layers it is possible to modify, update or replace them individually. Furthermore, this allows for the setting of certain requirements before proceeding to the next shell in terms of humanizing the system itself.

This could be managed using continuous verification, only allowing further proceeding after certain humanizing requirements have been met to a certain degree. Making this process iterative with intervals between updates, would allow for the combination of continuous verification and continuous compliance, only requiring the reverification of the layer, where compliance updates have to be made [1].

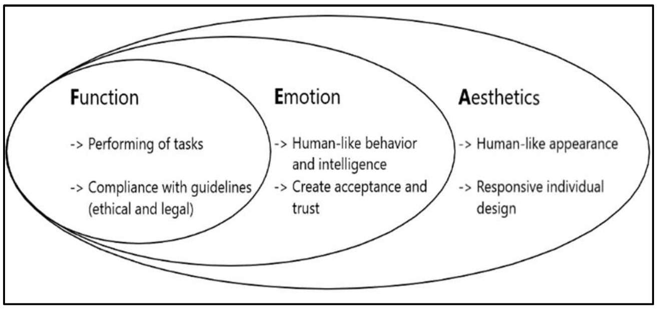
Function
-> Performing of tasks
-> Compliance with guidelines (ethical and legal)
Emotion
-> Human-like behavior and intelligence
-> Create acceptance and trust
Aesthetics
-> Human-like appearance
-> Responsive individual design

Fig. 1: A holistic view of the FEA Modell and its components

III. THE FUNCTION LAYER OF FEA

To be able to humanize A.I. certain requirements have to be met when producing the software itself. Brent Mittelstadts draws a comparison between policies in A.I. and medicine which resemble each other to a significant degree.

The major difference can be found in their goals. While diverging in its application to the real world, medicine has a fundamental consensus on what it is trying to accomplish namely the healing and supporting of patients. A.I. on the other hand, which is mostly used in the private economic sector, sets the focus mostly on maximizing the output with minimal cost [2].

A. Ethical Guidelines inside the Technical Layer

To tackle this problem of no set ethical standards on how A.I. should be used to achieve goals, the High-Level Expert Group on A.I., a special committee of experts appointed by the European Union, presented the “Ethics Guidelines for Trustworthy Artificial Intelligence”. According to the published guidelines these characteristics should be:

1. Lawful – respecting all applicable laws and regulations [3].
2. Ethical – respecting ethical principles and values [3].
3. Robust – both from a technical perspective while taking into account its social environment [3].

The FEA concept proposes to embed these guidelines into the ethically aligned design published by IEEE. Instead of integrating the eight general principles of the IEEE Design into EAD Pillars, we categories the eight general principles into the three categories proposed by the High-Level Expert Group, resulting in measurable pillars, made up from well split data, highlighting individual problematic areas (Fig. 2)[3][4].

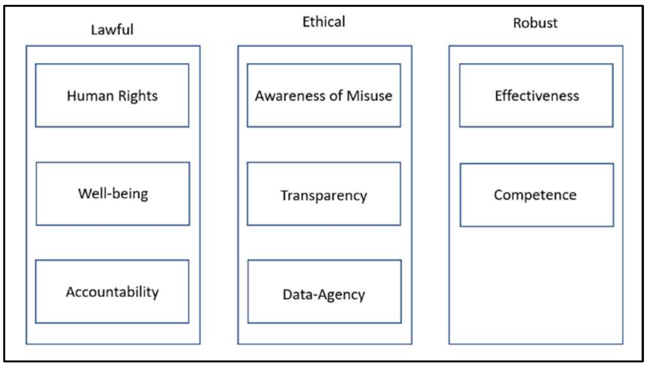

Fig. 2: Holistic view of the EAD Pillars fused with the 8 general principles of the IEEE ethically aligned design

As seen in in Fig. 2 this allows for a clear differentiation between characteristics and their assigned pillar. The advantage of this approach is its clear providence and transparency in which layer which components are present. The eight general principles are:

1. Human Rights – A/IS shall be created and operated to respect, promote and protect internationally recognized human rights [4].
2. Well-being – A/IS creators shall adopt increased human well-being as a primary success criterion for development [4].
3. Data Agency – A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people´s capacity to have control over their identity [4].
4. Effectiveness – A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS [4].
5. Transparency – The basis of a particular A/IS decision should always be discoverable [4].
6. Accountability – A/IS shall be created and operated to provide an unambiguous rationale for all decisions made [4].
7. Awareness of Misuse – A/IS creators shall guard against all potential misuse and risks of A/IS in operation [4].
8. Competence – A/IS creators shall specify, and operators shall adhere to the knowledge and skill required for safe and effective operation [4].

If certain ethical requirements inside one of the pillars should not be met, only the affected pillar needs to be examined. Inside the affected pillar the clear structure of division into these different characteristics allows for a clear limitation in which space an existing error is occurring. Also, this structure allows to set requirements not only for a certain pillar, but specifically for the principles it is made out of.

IV. THE EMOTION LAYER OF FEA

While the technical layer of FEA serves to accomplish the primary task of the product, the emotion layer is primarily concerned with analyzing the current situation of the user and respond in an appropriate manner. To achieve relevant insights FEA uses two main factors, namely the current emotional state and the trust level in the product.

The trust level can be dynamically determined using the quantitative trust model suggested by Hu et al. This model uses different criterions like gender and nationality to determine the trust level of a greater populace in an algorithm or product and reaches an accuracy of 92% for the general population [5]. Using this model with modified criterions like for example the age of the average user the product can be programmed to react differently with different users.

It should be noted though that some criterions like nationality are not easily detectable and should be either avoided or determined through other measures (e.g. using the location to determine a probable nationality). Additionally, feedback from the users is required which should be collected in a non-intrusive but continuous manner.

This may also serve as an indicator how actions done by the product affect the overall trust the user has in it. Through this continuous trust can be monitored. Furthermore, the A.I.

behind the emotion layer has the ability to learn from interactions and continuously improve itself.

The emotional state can be determined using an additional artificial neural network. A usable constellation for this purpose was proposed by Riaz et al. It uses facial recognition to determine one of seven emotions, i.e. anger, disgust, fear, neutral, sadness and surprise. It achieved an accuracy of 98.8% on the SAVEE dataset [6].

After the product identified an emotion it can react in an appropriate manner, for example by adapting the way information are transmitted, initiating a conversation, the use of certain phrases or other predefined measures. To determine an appropriate manner in which to convey information the information must first be categorized. For this FEA uses five categories:

1. Excellent
2. Good
3. Neutral
4. Bad
5. Awful

The overlying A.I. of the emotional layer, which is trained to mimic human speech patterns including platitudes, common phrases and other expressions, is now capable of forming a message which conveys the necessary information in an appropriate manner to the user.

This leads to a stacked software architecture in the emotion layer as can be seen in Fig. 3.

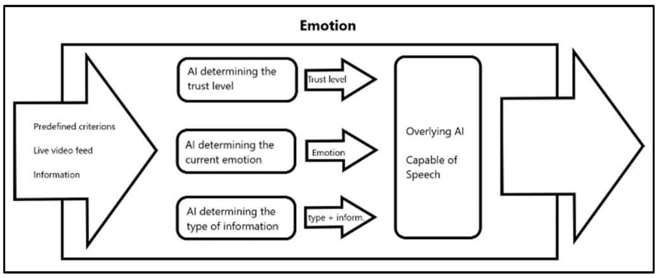

Fig. 3: Architecture inside the emotion layer

V. The aesthic layer of FEA

With the user interface being the part, with which the user is interacting, designing it to fit the user´s needs, individuality and preference is one of the most important aspects. Since the user´s perception, interests as well as preferences change over a longer span of time, continuous evolution of the user interface is a key factor to keep usage at a high over a significant timespan.
The approach here, would be to take the approach of universal design, modify it with A.I., which learns about the user´s preference more and more as time goes on. Universal design in its simplest form is the desire to make a design and composition of an environment so that it can be accessed, understood and used to the greatest possible extend by all people regardless of their age, size, ability and disability [7].

Not only having a basic understanding of intuitive design and the users´ needs but understanding what a user expects a certain function to do, as well as understanding the emotional response a user has to this function, is fundamental to good design. Taking the approach to create a basic design, after the ISO standard proposed in ISO 9421 and iteratively modifying it using A.I. gives a good basis to start from, increasing the tailoring for the user as time goes on [8].

A second important role could be assigned to the idea of aesthetic intelligence - making use of for example automatic assessment of image aesthetics allows for creative recommendations, photo ranking and personal album creations to show to the user. As can be seen in Fig. 4 the aesthetical value varies based on the user´s personal perception of the image [9].

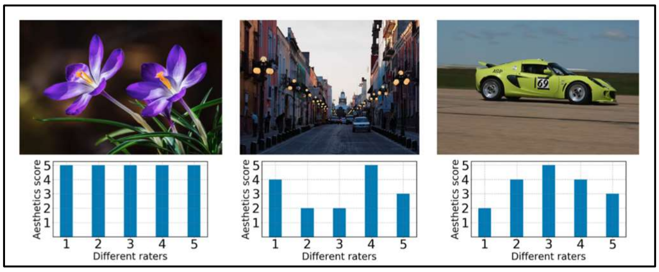

Fig. 4: Subjective aesthetical picture comparison on a scale from 1 to 5 as rated by 5 different users [9]

Since image aesthetics are highly subjective personalized imagine aesthetics aims to address the highly subjective factor and models itself after the user´s preferences. It therefore shares the overall goal of the aesthetics module.
This process can be split into the personalized prediction and the active learning part [9].

A. Personalized Predicition

Making use of collaborate filtering has been a popular and highly recommended algorithm for learning personal preferences based on a historical approach. It assumes that in principle a user who rates items high, will continue to do so in the future. In earlier works this has been done by adjusting weights of features in an ad-hoc way rather than in learning from data [9].

B. Active Learning

Active learning describes a framework in which initially available data has no labels assigned. Therefore, there is no existing reference between the input x and an associable label y. The goal is therefore to allow any label to access all initial data and adapt to the results of previous label requests [10]. This example shows one of many ways how an A.I. could adjust its look and its aesthetical factors to the user’s needs, interests and preferences.

VI. Studies

There exist several studies proving an increase in acceptance through the humanization of robots.

One such study was executed by Kupferberg et al. which concludes from a multitude of other studies that the brain processes biological and non-biological movements in a different way and that its perception plays a broad role in social interactions. Its goal is to examine whether a biological velocity profile or a variability in the movement trajectory are necessary to trigger motor interference.
This is a phenomenon where an observation of another person's incongruent movement leads to a higher variance in one's own movement trajectory. Motor interference in robots is necessary for the perception of them as humanized interaction partners [11].

To examine this, the study applied a quasi-biological minimum-jerk velocity profile to the motion of a robot. The robot was then observed by testers performing congruent and incongruent movements. The results of this suggest that a robot moving with quasi-biological velocity may result in the same type of implicit perceptual processes as if performed by a real human.
Additionally, the study suggests that detailed facial features may compensate for a less realistic robot body and vice versa [11].

Another study by Iwamura et al. looked at the different acceptance levels between robots used as "tools" and robots used as "partners". "Tools" offer for example physical assistance and perform only their main task. "Partners" on the other hand are expected to provide more than that primary service. They should provide companionship, initiate interactions and other services on its own to contribute to enhance the friendly relationships [12].

The study examined the difference in acceptance level through the example of robots for shopping assistance that help elderly people during a normal shopping trip in a supermarket. It tested four different constellations [12].

1. Humanoid robot without conversation
2. Humanoid robot with conversation
3. Cart robot without conversation
4. Cart robot with conversation

The results of the study suggest that humanoid robots and an enabled conversation function lead to better social acceptance than a solely practical robot or no conversation function respectively. Most participants felt that through the conversation they are doing things with someone and/or perceived positive feelings.
The humanoid robot made most participants feel like doing things with someone.
The study concludes that both factors, used to humanize the robot, lead to an improvement of enjoyment and therefore the intention to use while not affecting the perceived ease of use [12].

VII. Conclusion

The lack of trust in modern technologies, especially when in connection with A.I., can be counteracted through humanization as proven by multiple studies. The here presented FEA concept provides a framework to achieve this humanization. Through the suggested layered approach, a separate view on each functionality is possible enabling the near complete separation of the individual technologies used to achieve the overall desired result.

The first layer executes calculations based on modern ethical and legal guidelines for A.I. systems, while the second layer emotionalizes the information dynamically in response to the user. The third layer serves solely to present the emotionalized information in a for the user pleasant way while also serving to create an appealing look. When the information is presented to the user it has reached its most abstract form.

Furthermore, the suggest concept provides the possible integration of multiple aspects of continuous engineering, i.e. continuous verification, continuous compliance, continuous evolution, continuous use, continuous trust and continuous improvement. This is on one hand achieved through its architecture and on the other hand by using layered A.I. systems and its overall purpose of increasing the trust in A.I. technologies.

As FEA is a new theoretical framework a study concerning its effectiveness and feasibility may reap some additional insight.

References

[1] B. Fitzgerald and K. S. Stol, "Continous Software Engineering: A Roadmap and Agenda" Article in Journal of System and Software, pp. 182-184, July 2015.

[2] B. Mittelstadt, "AI Ethics – Too Principled to Fail?", Paper Published in ArXiv 2019 DOI: 10.2139/ssrn.3391293, pp. 2-4, 2019.

[3] High-Level Expert Group on AI, "Ethics Guidelines for Trustworthy Artifical Intelligence", First Draft December 2018, Published Report / Study, pp.6-8, 8th April 2019.

[4] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https:// standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous/systems.html.

[5] Hu W., Akash K., Reid T. and Jain N., "Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions", IEEE Transactions on Human-Machine Systems, Vol. 49, No. 6, December 2019.

[6] Riaz H. and U. Akram, "Emotion Detection in Videos Using Non Sequential Deep Convolutional Neural Network", IEEE, 2018

[7] n.A., "Disability Act 2005" Part 6, 52.-Part II Chapter IA Interpretation 19A. 'universal design', pp. 45-46, 8th July 2005.

[8] International Organization for Standartization, "ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centered design for interactive systems", First Edition, Published 15th, pp.1-5, March 2010.

[9] Ren. J., Shen X., Lin Z., Měch R., Foran D. J., "Personalized Image Aesthetics", pp.1-2 Published in 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 October, 2017.

[10] Hsu D. J. "Algorithms for Active Learning", p.5, PhD thesis, Department of Computer Sciene and Engineering, School of Engineering, University of California, San Diego, 2010.

[11] A. Kupferberg, S. Glasaur, M. Hubert, M. Rickert, A. Knoll, T. Brandt "Biological Movement Increases Acceptance of Humanoid Robotos as Human Partners in Motor Interaction", pp.1-2 and 4-5, Published by Springer, London, 2009.

[12] Y. Iwamura, M. Shiomi, T. Ishiguro and N. Hagita, "Do Ederly People Prefer a Conversational Humanoid as a Shopping Assistant Partner in Supermarkt?", pp. 1-2. And 5-7, Research Gate, Lausanne, Switzerland, 2011.