Working Paper

AIX Design Framework with Character Development for Ethical AI

Sudha Jamthe, Charles Ikem
Stanford University, PolicyLab Africa

sujamthe@gmail.com

Abstract

Machines and Artificial Intelligence (AI) are increasingly becoming a part of human life from public to intimate spaces. A social robot or a chatbot AI interacts with humans building relationships in private situations. With AI Algorithms such as GPT-3 creating human conversations, the AI is getting to create the conversations learnt from the training data and will override the human written rules that give it personality. It is not clear who gives AI its personality that genderizes and humanizes it and there is no transparency on the ethical values of the AI. In our theory we proposed the Artificial Intelligence Design (AIX) framework with a foundational layer of character that captures fixed tenets of ethical values of the AI such as trust, transparency and fairness. This paper builds out the experiments to prove that an AI developed with personality and value of transparency is more trustworthy than an AI developed without personality or character.
The experiments were to prove that character with personality built into an AI helps to build ethical values of trust, transparency and fairness. Our preliminary results showed that voice recommenders with a personality like Alexa tend to be trusted more by people. So, personality built into AI makes humans trust the device more.

Introduction

To evaluate the AI using our AIX framework which is a design framework for human-computer interaction. Here, the character of the AI is the foundation layer. The 'character' tenets describe the ethical values of the AI and includes, ethical values of 1. Trust 2. Fairness and 3. Transparency. These character tenets translate to the next layer in the design which is "Behavior." This defines the behavior of the AI whether the AI is going to be assertive or jealous or flexible etc and builds into the personality of the AI. Today AI is designed without any character basis with random behaviors and displays a personality with no transparency on how it would behave in certain situations.

For example, below is a comparison of voice assistants from top tech companies conducted by UNESCO. Other than genderizing the AI, see how they described the AI's behavior by the companies as 'friendly', 'funny', 'humble' and 'spunky without being sharp' which are all behaviors of the AI Voice Assistant. We compared Google Home vs Alexa

in 2016 as a fun project and their tone of conversation was clear in how they tried to be funny and when we tried to have them talk to each other there was 'jealousy' displayed[1]. All this is done by engineers hard coding for some phrases without any basis for character or values. We believe that when AI becomes an integral part of our lives and does useful functions like senior care or engage with people in serious situations, their behavior has to be thoughtfully designed and grounded onto clear values defined as their character.

For example, Alexa reminds us to reorder items that we ordered previously on Amazon. It tracks our engagement and behaviors which could be used to enhance the Amazon shopping experience. Without the character tenets which ties to the value of transparency, it is dangerous for people to trust Alexa as a fun voice assistant if her role in the home is to collect more data to get us to buy more on Amazon. This is not transparent so we do not know the reasoning behind her behavior. We are here to take sides on what is right or wrong but for the AI to have a character foundation with values that ties to behaviors and personality of the AI so that users can make a decision to still buy and engage with an AI knowing that it shares a particular value that is important to them.

The purpose of this study is to conduct experiments to collect data on ethics in designing and the user interaction data to define the AI's character with tenets of trust, transparency and fairness. With this project we are proposing tech ethics experiments to build models and datasets to test the hypothesis that adding character with AI Ethical tenets for AI will ensure transparency and trust of AI with users.

1. Experiments on ethics of AI and how it impacts user's trust by comparing AI built without character development with ethical tenets that translates of trust, transparency and fairness for the AI to an AI voice assistant or bot that exists today that is built without character with ethical tenets.
2. Experiment on personality traits of AI for an existing AI out in the industry today to check AI's gender, humanizing features such as tone and inclusiveness in its engagement on AI confusion matrix of false positives to get empirical data to show character tenets that drive personality traits are more trustworthy and fairer to users.

[1] https://youtu.be/OLVf9TF1JXI

Theoretical background

AI is everywhere solving complex predictive, personalization, replication and handling multiple variables. Machine learning which is one of the widely used AI application is a function that maps a set of input to a set of outputs. The inputs are features such as weight/height while the outputs are predictions such as sex-male/female. From this we know that AI has but its own challenges such as understanding emotions, creating original content, biases like racial profiling and misgendering. As humans, we can relate to emotional nuances like body language, tone of voice and context. So, AI requires some form of human intelligence to work- so it learns from what is fed into it.
Traditionally, we design emotion and behavior through a product's tone of voice like-colors, product category and brand story. Then we moved to designing interactions and experiences. Now, we have to contend with designing user and product behavior powered by AI.
Think of the journey and output of creating an AI-enabled voice interface for customers or deploying a machine learning model to identify objects. We could ask? Where do buttons go? What colors to use? How do output data react to user input? What actions the input will bring?

So, what is AIXDesign? AIXDesign is the design of algorithms that determine the behavior of intelligent systems. In building the framework for AIX (Jamthe, 2020), the ideas were to develop a framework that integrates the design process and design activities (UX designer) as a process of understanding users and demystifying the needs of the input/output data and how this entire computational process is communicated at the interface. Modern design capabilities are also a function of task-oriented and information-oriented as is on the web (Garrett, 2009). Garrett's model, 'The elements of user experience' describes a digital product as a set of planes-surface, skeleton, structure, scope and strategy.
Expanding on this concept is to think about the application of human-centered design to AI from ideation to user research, prototyping and products that mirror the intended behavior of algorithms.

AIXDesign is about designing the system behavior from a human-centered perspective. The voice, visual UI, taptics, haptics, et cetera. The underlying question for us are; how do we shape the machine experience; how do we translate data into user interfaces and prototypes that are intuitive and useful?

The building blocks of AIX consists of 3 elements: *user needs, data needs and interface needs.*
User needs: describes user needs and behavior patterns via user research methods

Data needs: How to conceptualize and contextualize data to support user needs via I/O operations.
Interface needs: Describes how interfaces can enable user behavior and interaction with I/O.

User needs deals with how to model the behavior of the users and generating useful insights. Data needs is about discerning the nature and use of the data in modelling the information we get from the environment including users. How it is collected, processed, analyzed and communicated as output. Interface needs deals with what happens when the user pushes a button or speaks through the mic.
These building blocks requires different methods and tools to apply in practice. For example, user needs entails user research with tools to map the user journey such as storyboard, journey maps. Analysis of the product use and usability such as use case matrix and identifying latent needs such as ethnographic study and mental model evaluation. Data needs requires various tools for both qualitative and quantitative data collection, processing/pre-processing, analyzing and communication.

Interface needs requires finding a fit on the type of interface that will enable user behavior and how users will interact with the AI output. This involves workflow such as UI Design, wireframing, prototyping to describe the interaction path, communication and general system behavior.
We have built upon this to implement designing AI using five AIX elements that give personality to the AI. These are *gender, tone, communication style, actions and autonomy* and are aligned to the work of Carol Smith[2] (2020). These are developed as consistent design elements that define the AI and give it personality when incorporated in the AI's initial development.

[2] Smith, Carol (2020): Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution.
https://doi.org/10.1184/R1/12119847.v1

AIX Framework explained

Character and Personality to AI as Design Elements

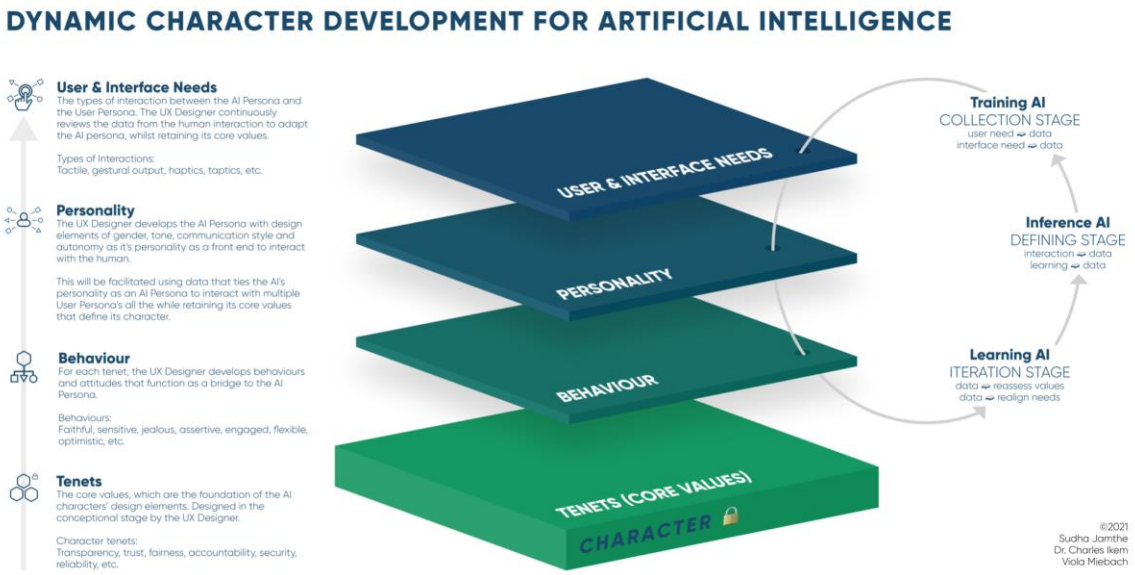


Fig 1 AIX Design Framework presented at EJEA at Kagawa by Sudha Jamthe, Charles Ikem and Viola Meibach[3]

AI personality, for a long time, was based on analogical reasoning in which information from a source is applied to a target through the connection of relationships or representations. By designing AI personality using UX, it becomes a context-dependent directive based on intuition, inferred knowledge or experiential understanding which increase the chances of reaching a satisfactory, but not necessarily optimal, solution. [1]

In the AIX framework, character is built as a foundation layer by UX designer to document and manage the character as part of the design specifications of the AI product. In this framework trust, transparency and fairness becomes fixed tenets of characters and this informs the behavior of the AI. From this is developed the personality of the AI which includes design elements of gender, tone, communication style and autonomy of the AI to be automated or agentive to ensure that the human computer interface is designed with agency to the human or AI thoughtfully by the UX Designer.

[3] Journal of Kagawa University with abstract of AIX framework presented https://www.kagawa-u.ac.jp/files/9316/5094/9800/Journal_of_KUIO_Vol.14_ver2.pdf

Character:

Gender: AI is a technology, and essentially a piece of software inference model that interacts with the user. The gender element gives a name to the AI and some elements that genderizes the AI. Today this is done with female names given to many AI providing support functions. UX design should consider Human Computer Design to decide if the AI needs to be gendered to humanize the AI's personality.

Tone: The tone of the AI can be designed to be formal and respectful or funny and light hearted or apologetic to create a personality that shows a certain age and relatable human behavior. For example, Amazon Alexa apologizes readily when she cannot find a requested song or misunderstands a user's ask but appears to get pushy when she recommends the user to add an item to buy proactively on Amazon by checking whether we are sure repeatedly.

Communication Style: In her work, Smith (2020) showed how to set up UX design with the right communication that is ethical. Prof. Ayanna Howard [12] has showed different ways for the AI or chatbot to help the user by being pre-corrective or post-corrective in helping a user in their choices by communicating proactively or to help only when the user fails and solicited or unsolicited to help the user when they ask for help or to proactively list options where the AI can help. This communication style is an AIX design element that should be thoughtfully designed and adapted to the growth of the relationship of the AI with the human.

Autonomy: This is a design element that tells how the AI acts in an agentive capacity to automate decisions for the human or acts in an assistive capacity like a chatbot would do to offer assistance to the human by receiving their input in an interactive fashion.

Methodology

We are proposing experiments to build models and datasets to test the hypothesis that adding character with AI Ethical tenets for AI will ensure transparency and trust with users.

1. We will take one existing AI in industry such as well-known chatbot or voice assistants (something that is engaging with customers today at home, car or public spaces, or online), and analyze its personality as we can see from its engagement. Then we run to A/B tests of two scenarios of the same AI, one assuming this personality was predictable and built using some character tenets and another that it was built without any character

tenet of trust, fairness and transparency. The experiments will be simulations of one to two of these consumer-facing AI.

We will run tests engaging these two AI on volunteers (similar to a survey) as a user research and collect data points of engagements and resulting trust of the user. We will collect enough of a sample set and build models to help derive consistent inference to show whether (and we hope it does) character tenets of trust and fairness help users in building trust on AI.

For example, if the AI is not built on a character tenet of trust, it might engage in saying one thing but might be collecting user data with a different intention and might break the user's trust at some point. In this case the experiment involves engaging with human subjects virtually for the user research.

2. The second experiment focuses on testing the hypothesis of earning user's trust with character tenets. For this one, we'll focus on fairness or bias tests against the personality traits seen in the AI. We will choose a set of AI publicly available and engaging with consumers. We will create an inclusive, diverse dataset of users. We will evaluate the personality traits of the AI's tone, style of conversation, gender and other personality traits for this dataset of users and record the AI confusion matrix to gauge its bias and fairness across a set of metrics. We will show how an AI that is built with a character tenet of trust and fairness earns the user's trust better from offering better accuracy and recall and serving the needs of a diverse set of users. This experiment is done by us with a dataset we develop to benchmark the AI's interaction in multiple scenarios. This will also involve humans but as virtual simulations run with invited volunteers including ourselves.

Hypothesis: The character development of Alexa with Transparency as a value tenet will enable users to trust the recommendations better.

Dataset: The authors of “Alexa, in you, I trust! Fairness and Interpretability Issues in E-commerce Search through Smart Speakers” made available the dataset which was created for their experiments. Using open source dataset collected by Dash et al (2022), Without running new models, we used their dataset and results to analyze patterns to draw a parallel for our AIX framework.

This dataset is built by comparing Alexa on the desktop with Alexa on the Amazon Echo device to compare the recommendations made for users. We see the desktop Alexa as an AI with no personality and the Alexa on the Amazon Echo device as having a personality and meets part of our AIX framework. We compare the results and stipulate

how adding a character tenet of transparency will improve user trust and hence prove our AIX Framework with character development is good to earn user's trust.

The dataset was collected in a manner where simultaneous query was placed with an automated voice for Alexa and Amazon desktop website to keep the geographical location constant. The first page of recommendations for a product search query was considered. Alexa has the configuration of selecting one product and adding it to cart. This method of recommendations of products is created by manufacturers and users have no power to change it.

Experiment 1

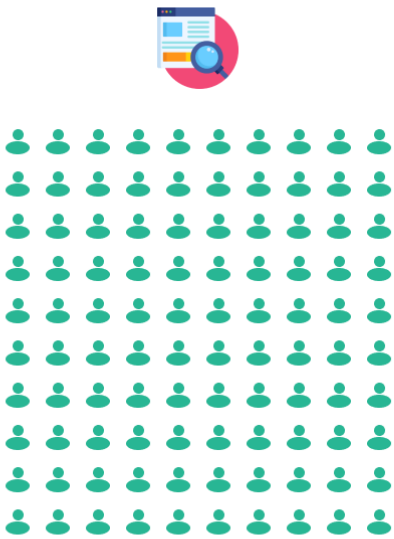
We analyzed the dataset and looked at the price difference of Alexa chosen products and the products recommended in the desktop search result.

The ranked recommendations in Amazon desktop results and the single product recommended by Alexa are analyzed for price differences. The ranked recommendations are limited only to the first page which implies they are the best recommendations.

Experiment Results:

We wanted to see if Alexa recommendations are any better than the desktop results and is there any advantage of buying from Alexa over a desktop website.

Working Paper



1000 People search for a product in www.amazon.com

The search results **average price of recommended products on desktop** will be compared to the recommended **products price in Alexa**

642 out of 1000
cases in the experiment,
the price of the product picked by Alexa is **more than the average price** of desktop results.

121 out of 1000 cases have a cost difference **greater than 1000 INR (Indian Rupees)** when compared to desktop results. It implies that Alexa is selecting costlier products which are not in the best interest of the user.

31 out of 1000 cases have a cost difference of **more than 10000 INR (Indian Rupees)** when compared to desktop results.

642 out of 1000 cases in the experiment, the price of the product picked by Alexa is more than the average price of desktop results.

121 out of 1000 cases have a cost difference greater than 1000 INR (Indian Rupees) when compared to desktop results. It implies that Alexa is selecting costlier products which are not in the best interest of the user.

31 out of 1000 cases have a cost difference of more than 10000 when compared to desktop results. This implies there is a certain intelligence given to Alexa to pick a higher price product which may give better quality.

According to the A. Dash et.al paper 40% are likely to buy and 10% are very likely to buy from Alexa recommendations.

Observation:

This shows that humanizing the voice, adding a character to Alexa gives rise to purchases although the recommendations are different from the desktop results. The character Alexa which has built trust needs transparency too. This layer of transparency when added gives the users an option to choose an affordable product.

Experiment 2

The products selected by Alexa were analyzed for any favoritism to particular characteristics of the products. Here we focused on the character tenet of fairness and how it will improve user trust of Alexa as an AI.

Experiment Results:

903 out of 1000 instances the products had the tag “Fulfilled by Amazon”. This implies that a certain bias exists in Alexa AI where the users are deprived of choosing from those products which do not have a partnership with Amazon. These biases form the personality of Alexa AI but giving it a character laden with transparency improves user agency to a wide variety of products.

334 out of 1000 instances the product chosen by Alexa AI was from Vendor 1 (Cloudtail)

289 out of 1000 instances the product chosen by Alexa AI was from Vendor 2 (Appario)

Observation:

There is a clear favoritism to particular vendors by Alexa. This is detrimental to healthy competition between vendors. There are chances that these may not be the best products and may result in dis-satisfaction from the customers.

Also, there is a clear favoritism for those products which are “Fulfilled by Amazon”.

In Spite of these favoritisms the Alexa character imbibes a false sense of trust due to lack of character development missing from the design.

Here the character of Alexa is biased towards vendors who have a partnership with Amazon. If the users get to know about this bias, they will lose trust in the whole concept of Alexa and not just any recommendations given by it. There is also a chance that users might stop usage of other services as well. This is discouraging to the Alexa ecosystem as the causal effect of these recommendations on the trust factor of the usage of Alexa is important issue to be studied,

Experiment 3:

The products selected by Alexa are analyzed for the number of ratings given by other users which is a major factor for online shoppers.

Experiment Results:

38 out of 1000 recommendation instances had less than 10 number of ratings

137 out of 1000 recommendation instances had less than 100 number of ratings

Observation:

Alexa AI appears to favor products of this condition leaving behind some products with less user rating. This bias will likely impact users who may prefer these products but do not know about them. So users may lose trust if they cannot afford or do not find the recommended products useful to them. Our AIX framework, if applied beyond personality to add character to add the value of transparency of Alexa's recommendation choice, will allow users to see transparently how Alexa makes its recommendations and will increase user trust in the long run.

Recommendation & Conclusion

People trust Alexa AI and continue to trust Alexa AI. The character of Alexa is built on the premise it offers the best product recommendation to the user. The selling point of Alexa AI is its customer first mindset. Nevertheless, Alexa AI appears to favor products of certain conditions leaving behind people of certain vendors. This bias will likely impact Alexa users as there is lack of transparency which might lead to user dissatisfaction. The application of the AIX framework when applied beyond personality to add character of transparency of Alexa's recommendation will improve trust amongst the users.

The Autonomous aspect of Alexa is to be amalgamated with Transparency to assure bias mitigation regarding vendors and the partnerships with them. Humans should be given greater agency of choice by providing information on a ranked list. This could be difficult in a Voice Assistant ecosystem but a way has to be found out to give agency to the human to select the product after comparison to other products.

Alexa AI can try to understand the persona of the customer and provide more personalized recommendations which may enhance both the user experience for the human using Alexa as well as for the vendors selling their products.

Further research is proposed to test this concept using quantitative approaches to separate the data used to train the AI as two streams to be used by the data scientist and the user's evolving relationship with the AI to be used by the UX designer. This will then be validated using a qualitative approach by defining the AI's core values and personality in the design process to test that the personality can be adapted while keeping the character fixed using data from industry.

Limitations of our research:

We declare that:

1. We were able to prove that voice assistants with a personality like Alexa versus AI assistants on a desktop interface tend to be trusted more by people. So, personality built into AI makes humans trust the device more.
2. Currently, no AI with a personality has core values as our model description, so it is understandable that the AI misuses the trust to sell and recommend more products without taking into account the best interests of the human user.
3. By building character as the foundation in which personality is built upon, we believe the AI will act in the best interests of its human owner than its human manufacturer. But this is not currently the case.
4. Limited dataset does not clearly reflect and able to describe various scenarios of interactions with Alexa voice assistant. This limits the generalizability of our study as more robust study is proposed in the future.

We acknowledge these limitations in our study while we believe that a more nuanced and transdisciplinary understanding of ‘character’ and ‘personality’ as described in studies on psychology is needed to further juxtapose the realities of everyday human-machine interaction for AI.

Acknowledgment

This research was supported with a grant from Notre-Dame IBM AI Ethics Lab

Bibliography

1. Jamthe S., Ikem C., Miebach V (2022) Character Development for Artificial Intelligence. Special Issue of EJEA Conference 2021 in Kagawa of the Journal of Kagawa University International Office. Vol 14. (201-202), ISSN 1884−8745. Available: https://www.kagawa-u.ac.jp/files/3316/4991/1809/14.pdf

2. Dataset for ““Alexa, in you, I trust! Fairness and Interpretability Issues in E-commerce Search through Smart Speakers” https://doi.org/10.48550/arXiv.2202.03934

3. US Defense Department Bhttps://basicresearch.defense.gov/Portals/61/Future%20Directions%20in%20Human%20Machine%20Teaming%20Workshop%20report%20%20%28for%20public%20release%29.pdf

4. Kevin T. Wynne & Joseph B. Lyons (2018) An integrative model of autonomous agent teammate-likeness, Theoretical Issues in Ergonomics Science, 19:3, 353-374, DOI: 10.1080/1463922X.2016.1260181

5. Trust in Human-Robot Interaction (Kindle Edition) by Nam, Chang S., Lyons, Joseph B.

6. https://www.nbcnews.com/mach/science/robot-manipulates-humans-creepy-new-experiment-should-we-be-worried-ncna900361 (Humans refuse to turn off robots and get attached to them)

7. Google Duplex demo at Google IO 2018 https://www.youtube.com/watch?v=D5VN56jQMWM

8. Nachmani, E, Adi, Y & Wolf, L. (2020). Voice separation with an unknown number of multiple speakers.*37th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research,* 119:7164-7175. Available at: http://proceedings.mlr.press/v119/nachmani20a.html .

9. In the book "Designing with Data*"* by Rochelle King, Elizabeth Churchill, and Caitlin Tan, the authors present a layered model of "data-driven," "data-informed," and "data-aware" design.

10. Ethically Aligned Design from IEEE Advanced Technology for Humanity by The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

11. Horstmann AC, Bock N, Linhuber E, Szczuka JM, Straßmann C, Krämer NC (2018) Do a robot's social skills and its objection discourage interactants from switching the robot off? PLoS ONE 13(7): e0201581. https://doi.org/10.1371/journal.pone.0201581

12. Smith, Carol (2020): Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution. https://doi.org/10.1184/R1/12119847.v1

13. ACM SIGCHI: HRI 2020 Keynote: Ayanna Howard: Are We Trusting AI Too Much? Examining HRI in the Real World. https://doi.org/10.1145/3319502.3374842

14. 5 elements of user experience by Jessie James (founder of AJAX)

15. Carney, Michelle: Building ML Products for people *https://medium.com/aixdesign/building-ml-products-for-people-d46ba5901031* Building ML Products for people. Learn from MLUX founder Michelle… | by Avantika M | AIxDesign

16. Data Driven Personana Development https://www.researchgate.net/publication/200086136_Data-driven_persona_development

17. Data Driven Persona Development Jennifer Sullivan McGinnBrandeis University and Nalini P. Kotamraju of IT Univ of Copenhagen https://www.researchgate.net/publication/200086136Data-driven_persona_development

18. Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., & Gummadi, K. P. (2022). Alexa, in you, I trust! Fairness and Interpretability Issues in E-commerce Search through Smart Speakers. *arXiv preprint arXiv:2202.03934*.

19. “Breaking the Barriers of Humans and Machines” Techcrunch, Oct 2015 by Sudha Jamthe https://techcrunch.com/2015/10/16/breaking-the-barrier-of-humans-and-machines/

20. Garrett JJ (2009). Customer loyalty and the elements of user experience. In Design Thinking: Integrating Innovation, Customer Experience and Brand Value. Edited by Thomas Lockwood. -3rd ed. 22(251-257). Allworth press. ISBN-13:978-1-58115-668-3

21. Fu KK, Yang MC, Wood KL. (2016). Design Principles: Literature Review, Analysis, and Future Directions. Journal of Mechanical Design 138(10).

22. Vogler, C. (2007) The Writer's Journey. Mythic Structure for Writers. Studio City, CA: Michael Wiese Productions.