AIS SIGPRAG
Pre-ICIS Workshop 2019
"Values and Ethics
in the Digital Age"

Humane Anthropomorphic Agents: The Quest for the Outcome Measure

Position Paper

Elisabeth André, Sarah Bayer, Ivo Benke, Alexander Benlian, Nicholas Cummins, Henner Gimpel, Oliver Hinz, Kristian Kersting, Alexander Maedche, Max Muehlhaeuser, Jan Riemann, Bjoern Schuller, Klaus Weber

Abstract

Artificial intelligence has become an integral part of our daily lives. Today, we engage with intelligent agents at home, on the street, and at work. Rapid advances in technological capabilities make such intelligent agents increasingly human-like. Anthropomorphic agents are characterized by a high degree of socialness, intelligence, and efficiency. They afford many opportunities (e.g., convenience, availability, automation) yet also bring along potential negative impacts on human users, such as uninformed decision making, loss of control, or intransparency. Thus, anthropomorphic agents mark a new quality of human-computer interaction that should consider values and ethics in their design process and outcome. However, typical outcomes to measure the quality of an intelligent agent from a user-centric perspective are limited to accessibility, usability, or user experience. In this position paper, we argue that in the design of anthropomorphic agents, we need to go beyond established HCI measures and propose a new outcome measure called "humaneness".

Keywords: Anthropomorphism, Intelligent Agents, Humane Information Systems, Transparency, Human Autonomy

1 Introduction

Intelligent agents increasingly interfuse our lives. The growing availability and use of chatbots, robo-advisors, voice assistants, and avatars in augmented or virtual reality are but some examples. Advanced intelligent agents employ design features relating to, for example, visual appearance, speech synthesis, discourse structure, and reasoning, that in part indicate a human nature of the agent. One recent prominent example is Google Duplex, an AI system for accomplishing real-world tasks over the phone.[1] Human users tend to respond to such designs with anthropomorphism, that is, by attributing human-like features, behavior, emotions, characteristics, and attributes to the computer agents (Epley et al. 2007, Pfeuffer et al. 2019). We refer to intelligent agents that employ a human-like design using AI technologies as anthropomorphic socially interactive intelligent agents, or in an abbreviated form anthropomorphic agents.

We posit that these anthropomorphic agents mark a new quality of human-computer interaction (HCI) as compared to more traditional, less social, less interactive, less intelligent agents. This gives rise to revisiting and refining the overall aim of designing, operating, and running such anthropomorphic agents. Specifically, we posit that a focus on traditional HCI outcome measures such as accessibility, usability, or user experience (UX) is not sufficient. Rather, we should aim for humane anthropomorphic agents taking values and ethics under consideration – specifically, humane anthropomorphic agents should support human autonomy and transparency above and beyond established functional and non-functional outcomes.

The challenge can be approached from different perspectives. On the one side, governmental agencies may define and enforce specific requirements to be implemented in anthropomorphic agents. For example, the State of California in the United States recently passed a law pushing the disclosure of chatbots on websites.[2] Supranational initiatives led by the OECD and the European Commission aim to spur the societal discourse and support nations in developing pertinent regulations. On the other side, providers of anthropomorphic agents may proactively reflect the design and enforce specific ethical guidelines. For example, Microsoft recently proposed six ethical principles that should guide the design of AI-based systems.[3]

Among information systems (IS) scholars, the need to capture the effects of IS usage as an important driver for advancing design and management accordingly is well known. For example, DeLone and McLean (1992, 2003) synthesized and advanced the collective quest of IS scholars for the dependent outcome measures. However, the net benefits in the IS success model primarily focus on economic outcomes like costs, time savings and sales. Going beyond this rather instrumental perspective, value sensitive design is an approach to design technology that accounts for human values (Friedman 1997, Friedman, Kahn, Borning 2008). Mingers and Walsham (2010) started a thorough discourse and argued in favor of increased ethical considerations in the field of IS. Hassan et al. (2018) followed this endeavour and discussed the philosophical meaning and responsibility of IS research. In between, Myers and Venable (2014) concretized the overall need and proposed six ethical principles informing design-oriented research. To conclude, the refinement and extension of the set of dependent

[1] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[2] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

[3] https://www.microsoft.com/en-us/ai/our-approach-to-ai

outcome measures that we as a scholarly community should focus on has received increasing attention by the IS community. Examples include the "Bright ICT" initiative by the AIS that takes a positive stance on shaping the future, the resulting project of Lee et al. (2018), who proposed goals and principles for the "Bright Internet", and Gimpel and Schmied (2019) who identified the risks and side effects of digitalization including ethical challenges and societal issues. On the practical side, the movement of the Center for Humane Technology is committed to enforce humane technology and thus strengthening individual and collective well-being while using technology (Center for Humane Technology 2019).

Despite these promising approaches, we lack a coherent understanding of what the dependent outcome measures should be in order to successfully guide the design process and capture the impacts of information systems. To this end, this position paper reflects on anthropomorphic agents as an avantgarde class of IS to make the case for new dependent outcome measures for humane anthropomorphic agents: humaneness. With this, it aims to spur a discussion among IS scholars on the human-centric design objectives and dependent outcome measures we should be considering in our future work.

In the following, we first give a brief insight into the class of systems that we call anthropomorphic agents. Subsequently, we discuss the need for a new dependent concept from an HCI perspective. We leverage the recent work published by the European Commission (EC HLEG AI 2019) to shape trustworthy artificial intelligence as a first step towards conceptualizing "humaneness".

2 Anthropomorphic Agents

The rise of anthropomorphic agents is fueled by recent advances in artificial intelligence (AI), especially in machine learning (Jordan and Mitchell 2015). Machine learning replaces the complexity of writing algorithms, that cover every eventuality, with the complexity of finding the right general outline of an algorithm — in the form of, for example, a deep neural network (LeCun et al. 2015) — and then the processing of data. In doing so, it increases the efficiency and effectiveness of modeling cognitive, affective, and interactive skills of anthropomorphic agents. Machine learning allows anthropomorphic agents intelligently adapting to their human user in real-time. However, leveraging AI typically comes at a cost in terms of lower transparency, loss of control, and lack of trust by human users.

Anthropomorphism is the attribution of human-like physical or non-physical features, behavior, emotions, characteristics and attributes to a non-human agent or to an inanimate object (Epley et al. 2007). Anthropomorphism as a human innate tendency has been well documented for a long time in the history of humanity. Early drawings about 30,000 years ago depict animals with human-like bodies (Dalton 2003). The main goal of the projection of human-like attributes to non-human agents is to facilitate the understanding and explanation of the behavior and intentions of the non-human agents (Epley et al. 2007). In previous research on the "computers are social actors" paradigm, Nass and Moon (2000) found that humans tend to apply social heuristics for interactions with computers that are imbued with anthropomorphic cues (see Table 1).

Table 1: Key terminology on anthropomorphic agents

Concept	**Definition**	**Source**
Anthropomorphism	The attribution of human-like physical or non-physical features, behavior, emotions, characteristics, and attributes to a non-human agent or to an inanimate object.	Epley et al. (2007)
Anthropomorphic cue	A cue is any animate or inanimate feature that can be used by individuals to infer some meaning and as a guide for future actions. An anthropomorphic cue is a cue of an IT system indicating a human nature and, thus, triggering anthropomorphism.	Hauser (1996), Smith and Harper (2003), Feine et al. (2019)
Anthropomorphic design	A plan for arranging elements of a system in such a way that human nature is indicated, and, thus, anthropomorphism is triggered.	Own definition
Intelligent Agent	An agent is any virtual character or robot that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. An intelligent agent is "a system that perceives its environment and takes actions that maximize its chances of success".	Russell, Norvig (2016)
Socially Interactive Agent	An agent (i.e. a virtual character or a robot) that is capable of interacting with people and each other using social communicative behaviors common to human-human interaction.	International Conference on Autonomous Agents and MultiAgent Systems (AAMAS) 2018 Track on Socially Interactive Agents
Anthropomorphic Socially Interactive Intelligent Agent (for short: anthropomorphic agent)	Socially interactive intelligent agent having an anthropomorphic design by leveraging AI/ML techniques.	Own definition

The social interaction with the machines exposed a seemingly unnatural reaction towards the computers, which not only led to socially appropriate manners towards inanimate objects, such as politeness (Nass et al. 1999), but it also led to emotional and positive reactions towards computers (perception of high quality information and friendliness, conformance with the computer's information) (Nass et al. 1996; de Melo et al. 2014).

Since anthropomorphism constitutes an opportunity to engage with users of intelligent agents, developers increasingly apply anthropomorphic designs to give humans a familiar feeling with intelligent agents, because a natural and personal

connection is otherwise missing. However, the decision for an anthropomorphic design often does not follow theoretical foundations or empirical insights but rather follows rules of thumbs and "best practices" which does not necessarily lead to optimal outcomes. Anthropomorphic design invokes anthropomorphism on the side of the system's human users which makes it easier for humans to connect with the system and therefore facilitates the familiarization with it (Burgoon et al. 2000; Epley et al. 2007). The behavior on the part of the user can either be conscious or subconscious (see Kim and Sundar 2012) and both aspects are of importance for designing and understanding anthropomorphism.

Anthropomorphic design is multi-faceted and includes, for example, voice recognition as well as voice synthesizing and computer-graphical rendering of human-like faces or bodies, including mimics and gestures or wearable devices for experiencing a virtual reality. Anthropomorphism is about satisfying the visual or audible aspects of the interaction in a more human way as well as managing the content in that interaction (Pfeuffer et al. 2019). Previous work in natural language processing (NLP) focused on content. Recent advancements in machine learning and NLP techniques have enabled chatbots, for example, to behave more naturally and give more contextually sensitive responses, thus providing semantically correct answers and more trustworthy experiences in interactions, which in turn increases their human-likeness (Abdul-Kader and Woods 2015; Li et al. 2016) and likeability (Landwehr et al. 2011; Aggarwal & McGill 2007; Waytz et al., 2014). More readily available examples of such artefacts that have facets of anthropomorphic design are virtual assistants such as Google Home and Amazon's Alexa that can serve as some form of butler in a smart home. Recent advances in the area of social signal processing gave more emphasis to nonverbal and paralinguistic cues (Gebhard et al. 2018; Joo et al. 2019; Syed et al. 2018). Sophisticated socially interactive intelligent agents make use of multiple modalities (gestures, facial expressions, speech etc.) to converse with users and vice versa are also able to analyze such signals from the user.

One important goal of anthropomorphic design is to positively influence humans' affective responses when using intelligent agents, which is observed to be an important factor in human-computer interaction (HCI) (Hudlicka 2003; Ochs et al. 2017). Applying anthropomorphic design has proven to positively influence likeability up to a certain threshold level of human-likeness after which this effect however turns negative (Burleigh et al. 2013). This nonlinear effect makes the development of anthropomorphic agents particularly difficult and warrants further research that tries to understand and address these challenges successfully.

Researchers in specific areas have paid particular attention to the importance of anthropomorphic cues – examples are the field of robotics (Duffy 2003, Fong et al. 2003, Wiltshire et al. 2014, see also the proceedings of the International Conference on Social Robotics ICSR) and intelligent virtual agents like chatbots (Araujo 2018, Kim et al. 2018, Seeger et al. 2018, Benlian et al. 2019, Adam et al. 2019). Yet, there is no general theory, guideline or technology for anthropomorphic agents. At the same time, we observe in business practice the rise of different types of socially interactive intelligent agents with anthropomorphic design in various domains while developers are not fully aware of the impact of these designs on the effectiveness and dynamics of the interaction between human and system. Furthermore, by leveraging machine learning techniques it is becoming possible to adapt anthropomorphic design for

specific users, tasks and contexts. This comes with a lot of opportunities, but also threats with regards to possibly negative socio-economic outcomes for the user.

Figure 1 depicts an outline of anthropomorphic agents including multiple layers (input, agent core, output), the interaction with the human user and outcomes from such an interaction on various levels. The interactions we consider take place via one or more of the modalities text (visual verbal), speech (auditory verbal), image, video, or animation (visual non-verbal). We do not focus on auditory nonverbal modalities (music, environmental sound) as they are not central to anthropomorphic agents.

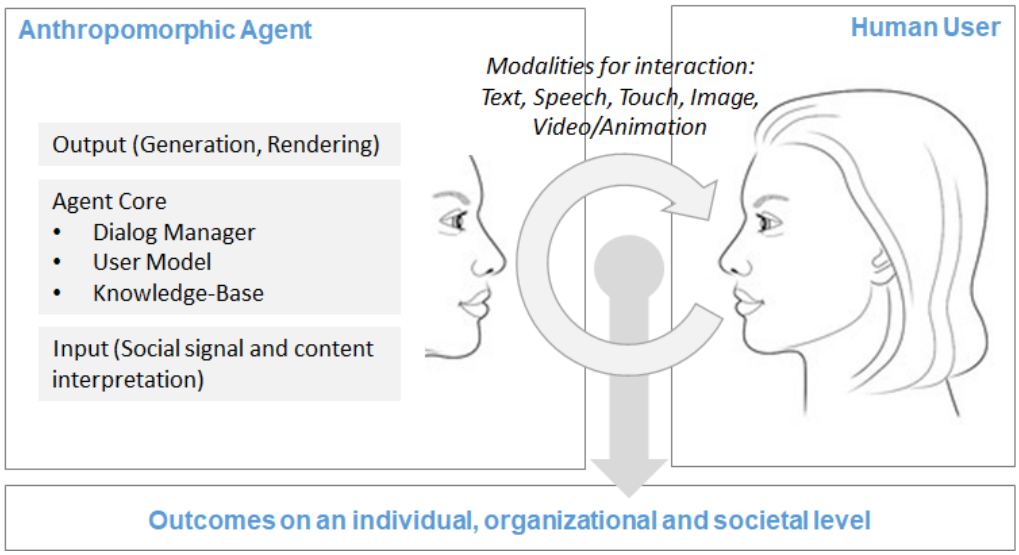

Figure 1: Anthropomorphic-Agent-User-Outcomes-Trifecta

3 Towards conceptualizing "humane" outcomes

3.1 Established outcome measures in HCI and IS

The HCI discipline has a long tradition of defining and analysing outcome measures of systems from a user-centric perspective. Specifically, accessibility, usability and user experience are important non-functional characteristics of interactive systems. HCI outcome measures are also recognized as important antecedents (e.g. perceived ease-of-use, effort expectancy, etc.) in well-known IS acceptance and use models, such as the Technology Acceptance Model (TAM), the Unified Theory of Acceptance and Use of Technology (UTAUT) and the IS Success Model. In parallel, a focus on economic outcomes (costs, savings, revenues etc.) frequently complements the assessment of net benefits of introducing or using a system.

In the following, we briefly introduce HCI outcome measures and their limitations with regards to capturing the "humane" perspective. First, accessibility (ISO 9241-171) is the extent to which a system enables users to interact with it, regardless of their level of vision, hearing, dexterity, cognition, physical mobility, etc. "Design for All" is a system design that considers human diversity, social inclusion and equality.[4] Several standards and guidelines for accessibility are available and have been legally enforced

[4] http://dfaeurope.eu/wp-content/uploads/2014/05/stockholm-declaration_english.pdf

in some markets. Relevant guidelines include W3C's Web Content Accessibility Guidelines (WCAG) 2.0 and ISO 9241-171 guidance on software accessibility. Accessibility emphasizes humaneness only from one specific viewpoint, namely non-discrimination of people with physical constraints or abilities.

Second, usability (ISO 9241-11) defines the extent to which a system, product, or service can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use. Usability is well established in the field and can be measured either subjectively by surveys (e.g. System Usability Scale, SUS) or objectively in usability tests. Furthermore, there exist well-established guidelines (e.g. Usability body-of-knowledge, https://www.usabilitybok.org/) to support the design of usable systems. One prominent example for usability engineering are Nielsen's 10 usability heuristics of user interface design (Nielsen, Molich 1990, Nielsen 1994). However, usability only captures user opinions about a system in the satisfaction dimension and does not emphasize any further human-centric perspectives. Nevertheless, the clear conceptualization and operationalization of usability positively impacted the way software and digital services are designed and delivered to its users in a sustainable way.

Finally, user experience (ISO 9241-210) describes user's perceptions and responses that result from the use and/or anticipated use of an interactive system. User Experience is a complex concept that is hard to operationalize and measure. For example, users' perceptions and responses include the users' emotions, beliefs, preferences, comfort, behaviors, and accomplishments that occur before, during and after use. Furthermore, user experience is a consequence of brand image, presentation, functionality, system performance, interactive behavior and assistive capabilities of a system. Despite its complexity, user experience has found its way into the real-world. Providers of software and digital services indeed successfully deliver "positive" experiences for their users. The underlying experience design process in many cases goes so far that users are manipulated with regards to the stimulated affective-cognitive states and the resulting user behavior.

3.2 Towards humaneness of anthropomorphic agents

Two examples of anthropomorphic agents shall illustrate that the measures introduced above are not sufficient. First, consider a voice-based assistance system that is so advanced in language processing, dialog management, speech synthesis etc. that in a phone call the human counterpart does not realize that she is talking to a computer. The system might be accessible and usable for most people and provide a great user experience in terms of positive emotions from a presumed human-human interaction. In fact, Google Duplex was "built to sound natural, to make the conversation experience comfortable [... so that] users and businesses have a good experience with this service"[5]. However, after the launch of Google Duplex intensive discussions among potential users with regards to transparency (specifically self-disclosure) were raised. This important humane outcome perspective is currently not covered by established HCI measures. Second, with regards to AI-based digital assistants in general, "persuasive systems can be supportive and engaging, but may lead to addiction [and] automated decisions (e.g., IoT devices ordering products) may be convenient, but deprive us of control" (Maedche et al. 2019, p. 540). Both these effects reduce human

[5] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

autonomy and might not be desirable, yet this effect is not covered by the above outcome measures. Thus, we argue that beyond the outcome measures introduced above, we explicitly need to capture values and ethics in new dedicated dependent outcome measures. Specifically, we suggest humaneness as one potential candidate concept to be considered. Figure 2 sketches the different individual-level-centric outcome measures in a nested model. As argued above, further outcome measures on the organizational and societal level may complement the outcome measures included in Figure 2. We limit our scope in this position paper on the individual level. However, there may be interesting trade-offs, for example between the degree of humaneness of an anthropomorphic agent and the targeted economic outcomes on the organizational level.

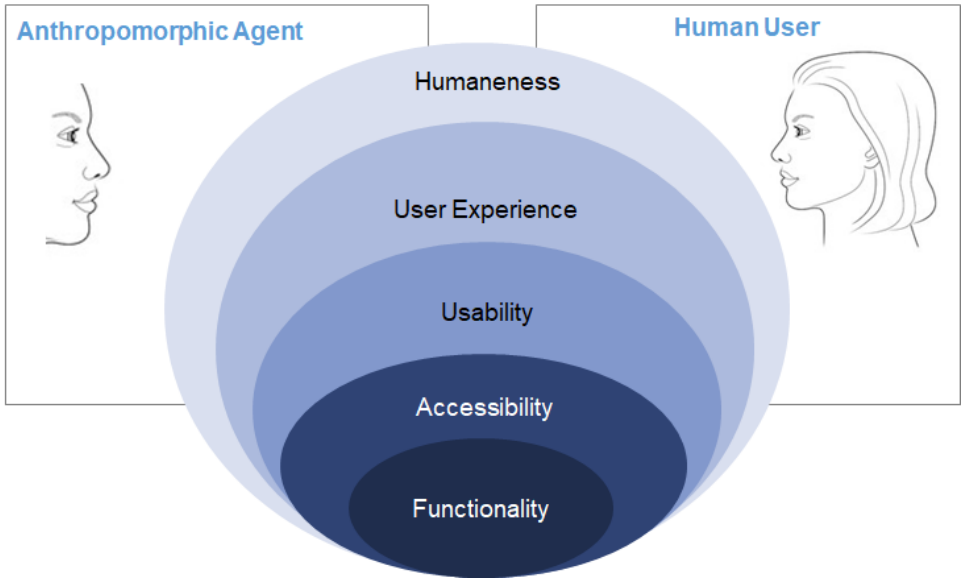

Figure 2: Nested model of potential individual-level outcomes

When humaneness might be a relevant outcome in the interaction of humans and anthropomorphic agent, the question is what exactly humeness comprises. A starting point is the recent discourse on the values and ethics of AI-based systems as broader class than anthropomorphic agent. From this starting point, we will identify the sub-dimensions of humaneness most closely related to the specific context of anthropomorphic agents.

Recently, the advances of AI-based systems spurred a discourse on their values and ethics. Two high profile supranational initiatives include the OECD principles on AI and the European Commission's ethics guidelines for trustworthy AI (see below for details). These are complemented by a discourse in academic literature (e.g., Bostrom and Yudkowsky 2014, Rahwan et al. 2019, Maedche et al. 2019, Fritz et al. 2019). The OECD assembled more than 50 experts from 20 governments as well as leaders from the business, labor, civil society, academic and science communities to develop not legally binding principles for AI-based systems. The recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI: "[1] AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being. [2] AI systems should be designed in a way

that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society. [3] There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them. [4] AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed. [5] Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles."[6] In a similar fashion and direction, the European Commission assembled a High-Level Expert Group on Artificial Intelligence (EC HLEG AI 2019). This group suggested that trustworthy AI should be lawful, ethical and robust. They identified four specifically pertinent ethical principles (respect for human autonomy, prevention of harm, fairness, explicability) and from this derived seven key requirements for trustworthy AI: (1) Human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental wellbeing, (7) accountability (EC HLEG AI 2019).

Both the OECD recommendation and the European guidelines suggest that the principles and requirements are complementary and should be considered in totality. Yet, they are not operationalized to a level that is applicable in the development and outcome measurement of anthropomorphic agents. Further, unlike for the recommendations and guidelines, our focus in this paper is not on AI per-se, but specifically on anthropomorphic agents leveraging AI, but adding anthropomorphism. We posit that for this specific class of systems, two outcome dimensions are especially relevant:

1. Human autonomy (related to European requirement #1, OECD principle #2).
2. Transparency (related to European requirement #4, OECD principle #3).

The focus on human autonomy results from the role of anthropomorphic agents as agents of their human users. An (anthropomorphic) agent acts on behalf of her/his (human) principle. When the agent stretches her/his duties too far, does not sufficiently understand her/his principle's values and preferences, or boundaries between what the agent decides and what the principle decides blur, human autonomy is at risk (Maedche et al. 2019, Fritz et al. 2019). The focus on transparency results from the anthropomorphic design of anthropomorphic agents. It blurs the ontological difference between human and machine and in part evokes affective and cognitive processes related to human-human interaction. In these blurring boundaries along with the novelty and complexity of artificial intelligence, transparency including traceability, explainability and communication appear paramount.

In suggesting this focus on human autonomy and transparency as specific outcome concepts to consider in the interaction between anthropomorphic agent and human, we deliberately accept and acknowledge that focusing on selected aspects goes along with not focusing on other aspects. This is unavoidable and partly intended. Other aspects put forward by the initiatives and academic discourse referenced above (e.g. non-discrimination or safety) are doubtlessly relevant for AI-based systems and, likely, for other classes of technical systems. However, in the specific context of anthropomorphic agents, we see that these two outcome measures stand out. We knowingly accept the

[6] https://www.oecd.org/going-digital/ai/principles/

narrowing in order to sharpen the discussion and stimulate operationalization. Here, we see an analogy with usability, a concept that found its way into practice of software development via focus and operationalization.

We conclude that a construct that concentrates on the humane outcomes of development and use of anthropomorphic agents is needed and that it needs to be multi-dimensional, covering specifically human autonomy and transparency. Following the structure suggested by the European Commission's High-Level Expert Group on Artificial Intelligence and again focusing on constructs most specifically related to the context of anthropomorphic agents, we suggest to further disaggregate humaneness as shown in Figure 3.

Figure 3: Disaggregated model of humaneness

Selected definitions of the constructs included in Figure 3 are provided in Table 2 as a starting point to develop a conceptualization for the context at hand. These specific constructs represent the core of the multi-dimensional construct humaneness, since they are experienced within the usage of a system.

"Humaneness" is the noun to the adjective "humane", meaning marked by compassion, sympathy, or consideration for humans users (based on Merriam-Webster online dictionary). This very broad concept can be applied to many technologies. In the 1960s and 70s, in light of destructive technologies like the hydrogen bomb, a research stream explored combining humanism with technology to produce humane technologies (e.g., Aspy 1975). In the 1990s, research on "Cognitive Technology" searched for characteristic features of humane interfaces in human-technology-interaction. This research was mainly concerned with the impact of technology on the mindset of the humans who apply them (e.g., Gorayska et al. 1999). Both of these research streams are to a certain degree related to humane anthropomorphic agents, yet they do not capture the specifics of this contemporary class of systems with their social interactivity, intelligence, and anthropomorphic design. Future research may review these research streams to identify potential elements supporting an operational handling of humane anthropomorphic agents. Based on the above considerations, we define humane anthropomorphic agents as *anthropomorphic socially-interactive agents that afford their human users' autonomy and transparency above and beyond established functional and non-functional outcomes*.

Table 2: Selected terminology for disaggregating humaneness in the context of anthropomorphic agents

Concept	**Selected Definitions**
Humaneness	Compassionate, sympathetic and designed with consideration for human users. (based on Merriam-Webster online dictionary)
Human autonomy	- Self-directing freedom and especially moral independence of humans. (based on Merriam-Webster online dictionary) - Individuals are free to make choices about their own lives, be it about their physical, emotional or mental wellbeing. (EC HLEG AI 2019) - Self-determination via the right to decide about being subject to systems' decision making or interacting with a system. (EC HLEG AI 2019) - Respect for self-determination and choice of individuals. (Beauchamp, Childress 2001) - Behavior that emanates from one's integrated sense of self. (Deci, Ryan 1995)
Human agency	- The human capacity, condition, or state of acting or of exerting power. (based on Merriam-Webster online dictionary) - System should support fundamental rights and the overall wellbeing of the user by supporting individuals in making better choices in accordance with their goals. (EC HLEG AI 2019) - Users should be able to make informed autonomous decisions using the system. (Bandura 1989) - Humans make causal contribution to their own motivation and action within a system of triadic reciprocal causation. (Elder. 1994) - Within the constraints of their world, people are planful and make choices among options that construct their life course. (Clausen 1993)
Human oversight	- Regulatory supervision of systems by humans. (based on Merriam-Webster online dictionary) - The capability for human intervention in either every decision cycle of the system, during the design cycles of the systems, or the overseeing of all activities of the system. (EC HLEG AI 2019) - Mechanisms such as the human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. (EC HLEG AI 2019)

Transparency	- The quality or state of being transparent, i.e. readily understood and free from pretense and deceit. (based on Merriam-Webster online dictionary) - Systems should be auditible, comprehensible and intelligible by human beings. (based on EC HLEG AI 2019) - The system should be traceable and should explain itself. (Mark, Kobsa 2005) - System transparency refers to when a user can immediately understand what the system is doing and how, by viewing the interface. (Rader et al. 2018) - Transparency involves encountering non-obvious information that is difficult for an individual to learn or experience directly, about how and why a system works the way it does and what this means for the system's outputs. (Rader et al. 2018)
Explainability	- Ability to give the reason for or cause of artefacts or behavior of a system. (based on Merriam-Webster online dictionary) - Capability to describe, inspect and reproduce the mechanisms and data used and produced by a system. (based on EC HLEG AI 2019) - Closely related to the concept of interpretability: systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation. (Biran, Cotton 2017) - The degree to which an observer can understand the cause of a decision. (Miller 2018)
Disclosure	- The act of making known the agent's artificial nature. (based on Merriam-Webster online dictionary) - The agent is identifiable as artifact rather than pretending to be human and communicates its capabilities and limitations. (based on EC HLEG AI 2019, called "communication" there)

4 Call to action

This position paper makes a case for expanding the set of outcome measures considered when assessing the aims and success of information systems. To constrain the scope of this considerable and complex endeavour, we focus on anthropomorphic agents. In this context, we posit that humanness is a relevant and appropriate target outcome complementing the traditional human-centred outcome measures accessibility, usability, user experience as well as further outcome measures on the organizational and societal level. We further suggest to disaggregate humaneness in six lower level, more specific constructs. Based on this first draft, we see the need for further research and a discourse among scholars and practitioners to refine and operationalize the conceptualization. Specifically, we would like to invite others in joining us in the following endeavours:

1. Validate the prioritization of specific constructs in disaggregating humaneness of anthropomorphic agents.

2. Synthesize and iteratively refine the definition of humaneness and its sub-constructs.
3. Operationalize the constructs in the form of survey scales to support the investigation of existing (humane) anthropomorphic agents.
4. Build exemplars of humane anthropomorphic agents to gain experience with their design and evaluation and derive specific design guidelines, design principles, and checklists.
5. Build a cumulative body of descriptive knowledge (Ω knowledge; Gregor and Hevner 2013) on the effect of anthropomorphic design on humaneness of anthropomorphic agents and on the interrelation and trade-offs with other, more traditional outcome measures (specifically economic outcome measures).
6. Build a cumulative body of prescriptive knowledge (Λ knowledge; Gregor and Hevner 2013) on methods for the engineering of humane anthropomorphic agents as well as generic design principles guiding the design of humane anthropomorphic agents.

Obviously, one might extend the study of humaneness beyond anthropomorphic agents to other classes of systems. This would likely include revisiting all the above steps. Both the perspective on descriptive and prescriptive knowledge are necessarily interlinked to support the development, diffusion, and use of humane anthropomorphic agents. We see this as a pluralistic, interdisciplinary research challenge that would benefit from contributions from information systems and computer science and a discourse with scholars from the social sciences engaged in ethics of technology, computers, and information.

References

Abdul-Kader, S. A., & Woods, J. C. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, *6*(7), 72-80.

Adam, M. T. P., Toutaoui, J., Pfeuffer, N., & Hinz, O. (2019). Investment decisions with robo-advisors: the role of anthropomorphism and personalized anchors in recommendations. In *Proceedings of the 27th European Conference on Information Systems (ECIS 2019)*.

Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer Research*, *34*(4), 468-479.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183-189.

Aspy, D. N. (1975). The humane implications of a humane technology. *Peabody Journal of Education*, *53*(1), 3-8.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*(9), 1175.

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press, USA.

Benlian, A., Klumpe, J., & Hinz, O. (2019). Mitigating the Intrusive Effects of Smart Home Assistants by using Anthropomorphic Design Features: A Multi-Method Investigation. *Information Systems Journal*, forthcoming.

Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8, p. 1).

Bostrom, N. & Yudkowsky E. (2014). The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence, 316-334.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human-computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior*, *16*(6), 553-574.

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*(3), 759-771.

Center for Humane Technology (2019). https://humanetech.com/

Clausen, J. A., (1993). *American Lives*. Free Press.

Dalton, R. (2003). Lion man takes pride of place as oldest statue. *Nature*, 425, 7.

Deci, E. L., & Ryan, R. M. (1995). Human autonomy. Efficacy, agency, and self-esteem. Springer.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60-95.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: a ten-year update. *Journal of Management Information Systems*, *19*(4), 9-30.

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(1), 73.

Duffy, B. R. (2003). Anthropomorphism and the social robots. *Robotics and Autonomous Systems, 42* (3-4), 177-190.

EC HLEG AI - European Commission, Independent High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*. Brussels, April 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Elder, G. H. (1994). Time, Human Agency, and Social Change: Perspectives on the Life Course. *Social Psychology Quarterly, 57*(1). 4-15.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies, 132*, 138-161.

Friedman, B. (Ed.) (1997). *Human Values and the Design of Computer Technology.* Cambridge University Press, New York, NY.

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. *The Handbook of Information and Computer Ethics*, 69-101.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems, 42*(3-4), 143-166.

Fritz, A., Brandt, W., & Gimpel, H. (2019). „Moral agency" ohne Verantwortung? Analyse von drei ethischen Verständnismodellen der Mensch-Maschine-Interaktion in Zeiten künstlicher Intelligenz. *Digital Humanity - Ethical Analyses and Responses in an Age of Transformation*, Annual Conference of Societas Ethica, Tutzing, June 2019.

Gebhard, P., Schneeberger, T., Baur, T., & André, E. (2018, July). MARSSI: Model of Appraisal, Regulation, and Social Signal Interpretation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* 497-506. International Foundation for Autonomous Agents and Multiagent Systems.

Gimpel, H., & Schmied, F. (2019). Risks and side effects of digitalization: a multi-level taxonomy of the adverse effects of using digital technologies and media. In *Proceedings of the 27th European Conference on Information Systems (ECIS 2019)*.

Gorayska, B., Marsh, J., & Mey, J. L. (1999). Methods and Practice in Cognitive Technology. In *Humane Interfaces: Questions of method and practice in Cognitive Technology*. North-Holland.

Gregor, S., & Hevner, A. R. (2013). Positioning and Preenting Design Science: Types of Knowledge in Design Science Research, *MIS Quarterly*, 37(2), 337–355.

Hassan, N. R., Mingers, J., & Stahl, B. (2018). Philosophy and information systems: where are we and where should we go?. *European Journal of Information Systems, 27*(3), 263-277.

Hauser, M. D. (1996). *The evolution of communication*. MIT Press.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies*, *59*(1-2), 1-32.

Joo, H., Simon, T., Cikara, M., & Sheikh, Y. (2019). Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in A Triadic Interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10873-10883.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless?. *Computers in Human Behavior*, *28*(1), 241-250.

Kim, S., Zhang, K.,, & Park, D. (2018). Don't want to look dumb? The role of theories of intelligence and human-like features in online help seeking. *Psychological Science*, *29*(2), 171-180.

Landwehr, J. R., McGill, A. L., & Herrmann, A. (2011). It's got the look: The effect of friendly and aggressive "facial" expressions on product liking and sales. *Journal of Marketing*, *75*(3), 132-146.

Lee, J. K., Cho, D., & Lim, G. G. (2018). Design and Validation of the Bright Internet. *Journal of the Association for Information Systems*, *19*(2), 3.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Li J., Monroe W., Ritter A., Galley M., Gao J., & Jurafsky D. (2016) Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, Austin, 1192-1202.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering, 61*(4), 535-544.

Mark, G. and A. Kobsa (2005). The Effects of Collaboration and System Transparency on CIVE Usage: An Empirical Study and Model. *Presence, 14*(1), 60-80.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Mingers, J., & Walsham, G. (2010). Toward ethical information systems: the contribution of discourse ethics. *MIS Quarterly*, *34*(4), 833-854.

Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management*, *51*(6), 801-809.

Nass C., Fogg B. J., & Moon Y. (1996). Can computers be teammates?. *International Journal of Human-Computer Studies, 45*(6), 669-678.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103.

Nass C., Moon Y., & Carney P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology, 29*(5), 1093-1109.

Nielsen, J., Molich R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* 249-256.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* 152-158.

Ochs, M., Pelachaud, C., & Mckeown, G. (2017). A User Perception-Based Approach to Create Smiling Embodied Conversational Agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(1), 4.

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic Information Systems. *Business & Information Systems Engineering*, *61*(4), 523-533.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson Education Limited.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2018). Designing anthropomorphic conversational agents: Development and empirical evaluation of a design framework. In *Proceedings of the 39th International Conference on Information Systems (ICIS 2018)*.

Smith, J. M., & Harper, D. (2003). *Animal signals*. Oxford University Press.

Syed, Z. S., Schroeter, J., Sidorov, K. A., & Marshall, A. D. (2018). Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech. *Interspeech*, 511-515.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 103. ACM.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477.

Wiltshire, T. J., Lobato, E., Velez, Jonathan, Jentsch, Florian, Fiore, & Stephen M. (2014). An interdisciplinary taxonomy of social cues and signals in the service of engineering robotic social intelligence. *Unmanned Systems Technology XVI*, 90840F.