How to Center AI on Humans[1]

Frank Dignum and **Virginia Dignum** [2]

Abstract. In this position paper we investigate what it means for AI to be human-centered. Although many organisations and researchers by now have given requirements for human-centeredness, such as: transparancy, respect for human autonomy, fairness and accountability, this does little to indicate how the AI techniques should be designed in order to be human-centered. In this paper we argue that human-centered AI involves a shift from AI emulating intelligent human tasks, to emulating human intelligence such that we capture enough social intelligence in order for the AI system to be able to center its activity and reasoning on its human users.

1 INTRODUCTION

In the past year many people in Europe have argued that research in AI in Europe should be human-centered. This would fit well with the European culture and distinguishes our research in AI from that in the USA and China. Although this sounds intuitively correct and governments and the European commission have embraced this perspective, little is known about what human-centered AI should look like. Is it enough to clad AI techniques in a social layer? E.g. by adding some natural language interface? The EU [14] gives a number of aspects that should be taken into account when developing AI systems in order to make them human-centered:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

These seem also quite reasonable requirements. However, if e.g. I develop a natural dialogue interface (which is clearly an AI system) to a service of my organization, which of these requirements apply? Let's just look at the fifth requirement.

We clearly should make this dialogue system respect diversity and be non-discriminatory and fair. But what does that mean? Address people based on their background to respect diversity? Or would this be discriminatory? And how would we define a *fair* dialogue? It is clear that these requirements are created mainly with a type of machine learning systems as AI system in mind. Systems that learn classifications from lots of data can make *unfair* decisions if a particular exceptional situation did not occur before, or did not occur often enough to warrant a correct decision. However, not all AI systems make decisions as their major outcome. Dialogue systems produce natural language based on the input of a user. Robots decide on autonomous behavior, which might be correct, efficient or stupid but not necessarily fair or unfair.

It seems we should not take the requirements as given by the EU (or other organizations) too literal, but rather as guidelines about the type of things that we should think about. Human-centered means that a system should have the human partner always as part of the focus for deliberation. This means that any task of the AI system should not be done in isolation, but the task should be done **for** someone, **in** some context (place and time). And if the actions of the AI system affect people directly or indirectly it should be aware of this and take it into consideration when deliberating. Thus e.g. if a system determines the best positions for windmills in a neighbourhood it should take into account the possible nuisance of the noise of these windmills for people living close by. Thus the AI system should be **socially aware**. In 1942, J. Gambs [8] defined being socially aware as:

> To know in every fibre of our body; to understand in its many ramifications and myriad applications the profound psychological principle that men and women have importance only as members of a group, that they can realize themselves only by giving themselves freely and generously to their group.

This quotation shows in more powerful words that being human centered means that everything one does should be for the benefit of the humans involved. As the quotation is about human social awareness it can talk about self realization which is one of the primary drivers of people. AI systems do not (necessarily) have this drive for self realization and thus dependence of the group of people interacting with it. However, this aspect can be emulated by the designers of the AI system by using a value based approach to create the system. I.e. using the values of the group for which the AI system is designed as the starting point to determine what it should strive for (what should its goals be or what it should optimize).

In this paper we argue that human-centered AI entails a paradigm shift in how AI techniques are developed and deployed. In the next section we discuss the specific social perspective that is needed. In section 3 we discuss more on how this can lead to genuinely human-centered AI. In section 4 we discuss how human-centered also means humanity centered and leads to what is nowadays is called ”AI for good”. We finish with some conclusions.

2 SOCIAL AI

The vision of human-centered AI, requires that AI systems are social. What does this mean and how to realise social AI is however a much less clear issue. Several authors, e.g. [13, 4], have argued that agents should become more aware of the social context in which they operate. This awareness is not included in the standard AI models of reasoning, such as the BDI model of agents, which focus on

[2] UmeåUniversity, Sweden, email: {dignum,virginia}@cs.umu.se

the goals and plans of an individual agent. What these authors argue for is a more social science based approach to the basic deliberation of AI systems. Although one can argue that this is not necessary in order to build an AI system that behaves as if it is social, it will make it a lot easier. Let us try to explain this more in depth.

If we talk about human-centered AI, we assume that the AI system's functions are directed and synchronized with the humans it interacts with. But how is this done? First we need to have at least some model of human behaviour that is good enough to predict what a human would expect from the AI system. This model can be fairly simple if the AI system is a mere classification or pattern recognition tool for the human. In these cases the only thing that one should know about the human is the optimization criteria that are used to determine the optimal decision of the human given the output of the system. E.g. if the system is used to determine whether a suspect of a crime should get out on bail or not, we should know what is the acceptable chance that such a person skips bail or commits a crime again. However, when the judge subsequentially wants to know how the AI system got to its classification and thus wants an explanation, the AI system should start functioning as a partner of the judge. Thus the explanation it gives should involve a more complex model of the judge. Is this a more conservative judge that would put the threshold for bail higher? Or is the judge someone that looks more in depth at the personal circumstances of the suspect and thus might feel that some input for the system is lacking? Based on a model of the judge the explanation should be geared towards one or the other element.

The above is still a simple example, but it is illustrative for the fact that maintaining a kind of BDI or utility based model of the human is not sufficient. Most decisions people make are not based on these kinds of rational models. People have basic values that drive their decisions, they relate to other people, which makes them sometimes follow the lead of someone else, they have personal needs and motives that they want to satisfy which influence their decisions as well and finally people keep to habits and practices just in order to keep life simple (see [12]).

If an AI system is human-centered it should interact appropriate with the human and thus have some awareness of these more complex (and social) aspects of human deliberation in order to support a user to achieve the right optimum.

In recent years, several researchers in both ABM and MAS, [13, 4, 15], recognise the need for new models of deliberation that bring together formalization and computational efficiency, with planning techniques, and expertise on empirical validation and on adapting and integrating social sciences theories into a unified set of assumptions [1]. In particular, these models need to describe how behaviour derives from both personal drives such as identities, emotions, motives, and personal values as well as from social sources such as social practices, norms, organizations [3]. Main characteristics of sociality-based reasoning are [5]:

- Ability to hold and deal with inconsistent beliefs for the sake of coherence with identity and cultural background.
- Ability to combine innate, designed, preferences with behaviour learned from observation of interactions. In fact, preferences are not only a cause for action but also a result of action, and can change significantly over time.
- Capability to combine reasoning and learning based on perceived situation. Action decisions are not only geared to the optimization of own wealth, but often motivated by altruism, justice, or by an attempt to prevent regret at a later stage.
- Pragmatic, context-based, reasoning capabilities. Often there is no need to further maximize once utility gets beyond some reasonably achievable threshold.
- Ability to pursue seemingly incompatible goals concurrently, e.g. a simultaneous aim for comfort and sustainability.

Our claim is that human-centered AI requires new types of architectures that are not primarily goal or utility driven, but are instead situation or (social) context based in order fulfil the above characteristics. In the architecture sketched in Figure 1 a first step into the direction of these social agents is given. The context management of the agent filters the (social) context to lead to standard behaviour appropriate for that context. Whenever the context is uncertain, not recognized or not standard a second process of deliberation is started based on the motives and values of the agent and the current concrete goals. After the performance of each behaviour there is a feedback loop that is used to adapt all the elements of the agent based on the rate of success or failure of the behaviour in that particular context. However, there is also an input to the context management from the internal drives of the agent. I.e. the agent will actively search for a context to satisfy some of its needs if it can. E.g. if one feels lonely then one will actively search for a situation in which one meets with friends and/or family. Thus context management is not just passively filtering the environment, but also directing focus on parts of a context or seeking it to get the right context. Sociality-based agents are fundamental to the new generations of intelligent devices, and interactive characters in smart environments. These agents need to be fundamentally pro-active, reactive and adaptive to their social context, because basically the social context with people is not a static given situation, but is actively created and maintained based on mutual satisfaction of motives, values and needs. Thus the agents not only must build (partial) social models about the humans they interact with, but also need to take social roles in a mixed human/digital reality and start co-creating the social reality in which they operate. More work is needed to test and validate social agent architectures such as the exemplary one suggested in Figure 1.

An interesting feature of the architecture in Figure 1 is that it is not just depicting a single AI system, but concerns the shaping of AI ecosystems comprising autonomous and collaborative, assistive technology in ways that express shared moral values and ethical and legal principles as expressed in e.g. binding codes such as universal human rights and national regulations. This requires the understanding, developing, and evaluating AI applications through the lense of an artificial autonomous system that interacts with others in a given environment.

It is important to be able to extend this line of research to understand and model the ethical dilemmas that arise from the need to combine multiple norms, preferences and interpretations, from different agents, cultures, and situations. In the next two sections we will discuss the consequences of a human-centered approach.

3 HUMAN-CENTERED AI

To understand the societal impact of AI one needs to realise that AI systems are more than just the sum of their software components. AI systems are fundamentally socio-technical, including the social context where it is developed, used, and acted upon, with its variety of stakeholders, institutions, cultures, norms and spaces. That is, it is fundamental to recognise that, when considering effects and the governance of AI technology, or the artefact that embeds that technology, the technical component cannot be separated from the socio-technical system (Dignum, 2019). This system includes people and

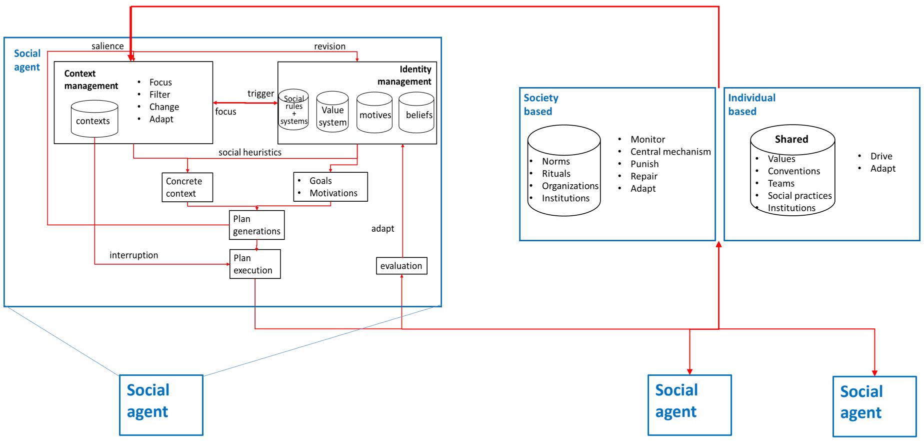

Figure 1. Sketch of a Social System Architecture

organisations in many different roles (e.g. developer, manufacturer, user, bystander, policymaker, etc), their interactions, and the procedures and processes that organise these interactions.

At the same time, it is as important to understand the properties of AI technology, as determined by the advances in computation techniques and data analytics. AI technology is an artefact, a software system (possibly embedded in hardware) designed by humans that, given a complex goal, are able to take a decision based on a process of perception, interpretation and reasoning based on data collected about that environment. In many case this process is considered 'autonomous' (by which it is meant that there may be limited need for human intervention after the setting of the goals), 'adaptive' (meaning that the system is able to update its behaviour to changes in the environment), and 'interactive' (given that it acts in a physical or digital dimension where people and other systems co-exist). Even though many AI systems currently only exhibit one of these properties, it is their combination that is at the basis of the current interest on and results of AI, and fuels public's fears and expectations [6].

Guidelines, principles and strategies must be directed to these socio-technical systems. It is not the AI artefact that is ethical, trustworthy, or responsible. Rather, it is the social component of the socio-technical system that can and should take responsibility and act in consideration of an ethical framework such that the overall system can be trusted by the society. The ethics of AI is not, as some may claim, a way to give machines some kind of 'responsibility' for their actions and decisions, and in the process, discharge people and organisations of their responsibility. On the contrary, AI ethics requires more responsibility and more accountability from the people and organisations involved: for the decisions and actions of the AI applications, and for their own decision of using AI on a given application context.

This also means that requirements for trustworthy AI, such as those discussed in the introduction, are necessary but not sufficient to develop human-centered AI. The *development* of human-centered AI systems should focus on more fundamental aspects of human responsibility such as values and norms. By starting from these fundamental social concepts the designers will be forced to define in terms of those concepts how they interpret the requirements as mentioned in the introduction. E.g. if "safety" is the primary value when developing the software of a self driving car, then the requirement of transparency might be interpreted as explaining why a certain action of the vehicle was safer than a default expected action. Thus transparency in this case would not include giving the whole causal chain of reasoning that led to the current action, but only that part that is relevant for safety. Moreover, there might be cases where a car producer does not want to give full transparency of the system as it could lead to exploitation of some particular preferences of the system with adverse effects. E.g. if it is known that any moving object that comes closer than 1.5 meter from the vehicle will cause the car to stop, people might use this to get right of way on the car preventing it from ever turning on a road.

From this example we can see two fundamental issues:

1. The AI techniques used in the AI system should be amenable to the ethical requirements such as transparency. I.e. it should be possible to explain (or to show) how the system got to a certain decision or behavior.
2. It should be possible to adjust the implementation of the requirement such as transparency based on the context in which the system is used. I.e. requirements such as transparency should not have one fixed definition for all AI systems, but rather be defined based on how the AI system is used.

The second statement seems to indicate that we could make any concrete definition of the requirements ourselves in a way that suits us best. However, this is not the intention. In order to make this more precise we could require that any concrete description of e.g. transparancy for a specific case should **counts-as** transparency in the sense as given by Grossi [10]. In this work the counts-as relation is defined such that when A counts-as B then A should at least contain the core of the meaning of B, but might have extra features in its penumbra. Thus one could state that a drivers licence (in some context) counts-as a valid ID, but club membership card (without a photo) would not counts-as an ID. The club membership card misses some of the core features. So, there is freedom in specifying what counts-as a concept, but not unlimited. In a similar vein one could state that the concrete implementation of the transparency requirement should be such that one can prove afterward that this implementation counts-as transparency.

We conclude that a truly human-centered AI system will exhibit such properties as emergent features from its design, but the mere adherence to these properties in a mechanical way does not make an AI system human-centered.

4 HUMANITY-CENTERED AI

Finally, in this context, it is important to discuss humanity-centeredness. In the previous section, we have mostly discussed the interaction with AI systems and its users, and how social awareness can improve this interaction and ensure trust in the system and its actions. Humanity can either mean an attitude, or moral sentiment of good-will towards fellow humans, or the collective existence of all humans [2]. Both definitions have been studied extensivly in psychology and the social sciences, which describe that humanity is necessary for our collective existence. However, the interests of individual humans and of humanity as a whole are not always aligned. In fact, individual solutions to shared problems may create a modern tragedy of the commons. For example, climatic changes, population growth, and economic scarcity create shared problems that can be tackled effectively through cooperation and coordination, but individual solutions to shared problems, such as privatized healthcare or retirement planning, can lead to inefficient resource allocations and coordination failure [9].

From an ethical perspective, the main issue often is the balance or dilemma between the good of the community and that of the individual. Social institutions are often the means to offer guidance in these aspects. The last few years have seen a proliferation of guidelines and principles for AI as a means to ensure that AI systems are designed and used both for the benefit of individuals and of society, [7, 11].

When AI systems are designed for humanity, requirements of inclusion, diversity, bias and well-being become leading. To serve humanity's best interests is the top priority of such AI systems, possibly leading to decisions that are less optimal for a given person or group. For example, AI systems aiming to solve the climate crisis may propose solutions that lower the living comfort levels that many are used to in the global North. Ways must be found for people around the world to come to common understandings and agreements - to join forces to facilitate the innovation of widely accepted approaches aimed at tackling wicked problems and maintaining control over complex human-digital networks. AI for Humanity is often equated with AI for Good, which promotes projects that have a positive impact on communities and humanitarian issues such as disaster management, agriculture, the environment, climate change, or promoting diversity and inclusion. As technology, AI has both the potential to contribute to solving or inhibit humanity's main challenges, as defined by the United Nations in the Sustainable Development Goals [16].

5 CONCLUSIONS

In this position paper we have argued that human-centered AI entails more than adding some social capabilities, such as explanation facilities, to AI systems. It is also not enough to give more precise or concrete definitions of concepts such as fairness and transparancy. The requirements for human-centered AI are the result of the combination of humans and AI system. Therefore, in order to have these properties emerge we cannot just impose some fairness condition on a system, but should design AI systems in a value based way, taking into account the social context in which the AI system is used. This also means that we have to have an eye for ethical dilemmas where optimality for humanity (or a larger group) can be different than for an individual. Making AI systems aware of their social context entails that they should be aware of the consequences of their actions for the humans they interact with. This means the AI systems should start using more realistic human models to predict expected behavior in the interactions. These models should at least incoprorate social concepts like social practices, norms, values, etc. Given this social context of human-centered AI it makes sense to develop AI systems that are themselves based on social deliberation mechanisms. We have provided a first sketch of how such systems might look. But, of course, much work needs to be done in this direction before thise type of systems can be fully utilized.

ACKNOWLEDGEMENTS

This work was partially supported by the Wallenberg Al, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] S. Chai, *Choosing an Identity: A General Model of Preference and Belief Formation*, University of Michigan Press, 2001.

[2] Robin M. Coupland, 'The humanity of humans: Philosophy, science, health, or rights?', *Health and Human Rights*, **7**(1), 159–166, (2003).

[3] F. Dignum, V. Dignum, R. Prada, and C.M. Jonker, 'A conceptual architecture for social deliberation in multi-agent organizations', *Multiagent and Grid Systems*, **11**(3), 147–166, (2015).

[4] F. Dignum, R. Prada, and G.J. Hofstede, 'From autistic to social agents', in *AAMAS 2014*, (May 2014).

[5] Virginia Dignum, 'Social agents: Bridging simulation and engineering', *Communications of the ACM*, **60**(11), (2017).

[6] Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer International Publishing, 2019.

[7] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al., 'Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations', *Minds and Machines*, **28**(4), 689–707, (2018).

[8] John S. Gambs, 'What does it mean to be socially aware?', *Childhood Education*, **19**(2), 51–51, (1942).

[9] Jörg Gross and Carsten K.W. De Dreu, 'Individual solutions to shared problems create a modern tragedy of the commons', *Science Advances*, **5**(4), (2019).

[10] D. Grossi, *Designing invisible handcuffs, formal investigation in institutions and organizations for multi-agent systems*, SIKS Dissertation series, Utrecht University, 2007.

[11] Anna Jobin, Marcello Ienca, and Effy Vayena, 'The global landscape of ai ethics guidelines', *Nature Machine Intelligence*, **1**(9), 389–399, (2019).

[12] D. Kahneman, *Thinking, fast and slow*, Farrar, Straus & Giroux, 2011.

[13] G. Kaminka, 'Curing robot autism: A challenge', in *AAMAS 2013*, pp. 801–804, (May 2013).

[14] EU-HLEG on AI, *Ethics Guidelines for Trustworthy AI*, 2019.

[15] B. Silverman, D. Pietrocola, B. Nye, N. Weyer, O. Osin, D. Johnson, and R. Weaver, 'Rich socio-cognitive agents for immersive training environments: case of nonkin village', *Journal of Autonomous Agents and Multi-Agent Systems*, **24**(2), 312–343, (March 2012).

[16] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini, 'The role of artificial intelligence in achieving the sustainable development goals', *Nature Communications*, **11**(1), 1–10, (2020).