Dignified Application of Artificial Intelligence: Advocating a Human Centered Approach

Marin Beijerbacht

December 2023

Abstract

Artificial Intelligence (AI) has become increasingly intertwined with our society. As such it has a large influence over our lives. Next to the advances AI has given us in handling large amounts of data it has had far reaching implications. Machine bias and loss of human autonomy are only two examples of how AI is hurting human dignity. In response to the problems with AI a human centered approach to AI has been proposed. In this article I will make a case for adopting this human centered approach to AI. Human Centered AI (HCAI) provides a framework that makes it possible to research, design and implement AI systems that align with human values and protect human dignity. I argue that this HCAI approach is needed to ensure the benefits of AI will outweigh it's harms and human dignity is protected.

Introduction

Artificial Intelligence (AI) has become increasingly intertwined with our society. From deciding which pieces of content are shown to a user for their enjoyment, to critical applications like advising judges and driving cars. Whilst AI has been beneficial in these and many other applications by for example taking some load off of humans in work places or helping innovation in research, not everything is rose coloured. Risks associated with the usage of AI, specifically in the machine learning area, are starting to emerge. Problems with, for example, accountability, explainability, control and fairness have already been identified in many applications. These can lead to new or exacerbate existing societal problems and cause harm for individuals [1,2]. In answer to these rising problems a framework has been proposed called Human Centered Artificial Intelligence (HCAI). It presents a way of thinking about AI technologies, their development and their deployment where protection of human values is at the center. It poses that instead of replacing humans with automated systems, AI should be created in ways that support the human and meaningful control of the technology is possible. One of the main proposals of the framework is that both high levels of human control and high levels of automation are possible, which will help produce systems that are reliable, safe and trustworthy [3, 4]. This all sounds promising, though one might ask, in order to advance the field of AI in a safe manner, is it necessary to adopt a human centred

approach to AI? In this paper I will argue that in order for the field of AI to advance in a safe manner a human centered approach should indeed be adopted, because it protects human dignity.

My argumentation is as follows. Currently, AI applications are hurting human dignity in different ways. The risk of hurting human dignity with AI will only increase in the future. From this it follows that something in the approach to AI has to change. A human centered approach to AI can provide the guidance needed in order to protect human dignity. Which together all leads to the argument that HCAI should be adopted for a safe advancement of the field of AI.

The next section will go into how human dignity can be conceptualised and is used in this paper. The next two sections present my full argumentation for the necessity of HCAI. The following section considers some objections to the thesis and replies to these after which the paper is concluded.

Defining Human Dignity

Before starting the argumentation, it is necessary to define how the term human dignity is conceptualised in this article. The definition of human dignity is somewhat ambiguous and has both a humanitarian and religious background. Dignity is most commonly associated with human rights discourse, where dignity is conceptualised along the lines of Kantian autonomy interpretation. Here dignity is a basic worth or status that all persons have. A human here is seen as an autonomous actor who can do moral reasoning, and who must not be instrumentalised as a means to an end [5]. Though through history and in its use in different types of articles, from human rights to bioethics it is argued to mean different things [6]. This ambiguity around its meaning has been used to argue that dignity is useless [7]. However, seeing human dignity as an essentially contested concept attacks this and shows why it can actually be a good thing [8]. An essentially contested concept can be seen as something "inevitably involving endless disputes about their proper uses on the part of their users", such as democracy or equality. To be a contested concept there are a few criteria, like having an agreed upon core meaning on which the other conceptions are based [9]. The different conceptions of human dignity center around an agreed upon core meaning, namely that human beings have an intrinsic worth and that this worth should be recognized and respected by others [10].

Because of all this, human dignity can also be applied in new contexts to extend existing rights and derive new ones [8, 10]. This is also the case for human dignity in the discussion around AI. In this paper I will use the conceptions of dignity as presented in [11]. In this paper four key conceptions distilled through case law and treaty interpretation are presented. Which due to the contested nature of human dignity are all equally valuable.

First there is the non-instrumentalisation conception, which has its roots in the Kantian interpretation of human dignity. Here an autonomous actor capable of moral reasoning, a human person, may not be used as a means to an end. Only with (informed) consent, such as working for a boss to generate profit, can a human being be instrumentalised. This does not hold for instances where the instrumentalisation of a consenting person would still categorically undermine the inherent worth of others.

The second conception is the protection of certain vulnerable classes of persons. Here

stronger protection is advocated for vulnerable classes of persons. Vulnerable in this context means that the person or group in question posses a socially salient characteristic that exposes to or facilitates exploitation or harm. Such as sexuality, race or having a disability for example.

The third conception is human dignity as the expression and recognition of self-worth. This self-worth is not something that is given, it is inherent in being human. Three forms of how self-worth is used in human rights are distinguished.

- First there is self-worth as respect to an individual as an autonomous being. Exercise of autonomy should be empowered, but also constrained when necessary. Exercise of autonomy is not completely individualistic, it can be relational when one engages social conditions. Hence, it is not implied that an individual is not allowed to have any interference in expressing their autonomy, it is bound by what is acceptable for them and the social structure around them.

- The next use of self-worth is used to express that harming someones worth by inflicting physical and mental mistreatment is prohibited. This also includes that denigrating or offending others is not accepted and harms their dignity.

- The third use of self-worth entails that some basic materials and conditions for a human person to at least survive and if possible thrive in. Here a duty to provide minimal social and economic goods can be derived.

The last conception of human dignity is the protection of humanity as a species concern. It is based on the idea that human beings are unique and therefore have a higher protection status. Some actions or technologies could bring harm to humans as a species. This conception has often been used in bioethics in the debate around altering the human genome [11].

The four conceptions presented here do not completely stand alone, as they do tend to overlap a little bit. For example non-instrumentalism and exercise of autonomy as linked to self-worth are both based upon a premise of human exceptionalism as these theories rest on only humans being able to do higher moral reasoning. They are split up in order to help make a mental model of the ways human dignity is used in practice. They are also not exhaustive as new ways of applying dignity can emerge, with new technologies.

Artificial Intelligence and Human Dignity

Hurting Human Dignity

The first premise of my argumentation is that AI applications and the business practices around them are hurting human dignity. To support this I shall go over a few specific AI systems that have already hurt the human dignity of individuals and discuss some wider traits of AI that are causing problems. These will not be exhaustive of all issues, though serve the purpose of illustrating how human dignity can be hurt by such systems.

The decisions that AI systems make and actions they take are based on what data it has been trained on, and what the designers have said the system should optimise. In the

data representing humans, physical phenomena and behavioural data are codified through, for example, clicks, typing speed or historical records about your health [12]. Through this AI does not handle with you as a person but with data representing you, whilst its decision can still heavily impact you. Add to this the bias that is existent in the data we use to train algorithms, the use of incomplete data or generalising too much. This all makes bias a common risk for AI [13]. Bias like this has been found in a widely used health care risk algorithm in the US used to prevent serious complications and reduce costs. Here black patients predicted at some risk level were sicker than white patients predicted at the same risk level. This happened as health costs were used as a marker for health needs, whilst the historical data the model was trained on contained bias [14]. This way the vulnerability as a class conception of human dignity is attacked, as well as the self-worth of the people in groups that get treated unfairly due to bias. When AI is applied to politically or monetarily driven ambitions such as catching welfare cheats or cutting costs, such as in the application above, human beings are easily instrumentalised. Engaging the non-instrumentalisation conception of human dignity when people are reduced to objects to be managed for other institutions goals and are reduced to data points without context [11].

Autonomy is an important concept for human dignity, capabilities for autonomous rational decision making are foundational to multiple conceptions [10]. One direct way in which AI could negatively impact autonomy is by taking away decisions for humans. However, there are less noticeable ways AI is damaging the autonomy of humans. In doing so attacks the human dignity as self-worth through respect for autonomy conception. Having AI systems embedded in services people use in their everyday lives shapes the choice architecture, the environment where decisions are made, by influencing the available options and choices. This shaping of the environment can influence cognitive autonomy to a point where people are manipulated. Manipulation here means that there is a hidden influence on another person's decision making, by unconsciously influencing emotions, behaviours, beliefs or desires [15]. Many AI systems are used in places where they are invisible and opaque to users whilst shaping their choice architecture. When used in the interest of an institution it can quickly become manipulative. One of the examples where this type of shaping was performed was uncovered on Facebook. Here the AI algorithm in use could detect when a user, specifically teenagers, felt insecure or worthless and served them ads based on this influential state [16]. This manipulation for profit attacks both the human dignity as respect for autonomy and non-instrumentalisation conceptions.

Threats to the human dignity as the protection of human species concern are not as abundant as threats to the other ones as discussed above. Though they do exist, most notable of which comes from autonomous weapons systems. By seeding control to machines for warfare there are fears that things will escalate. However, the most pressing current concern with these systems is still the non-instrumentalisation conception of human dignity. Combatant victims of autonomous weapons are said to be instrumentalised as 'simple objects of machine action' [17].

The Near Future

My second premise is that with wider adoption of AI, the harm done to human dignity will only increase. In the past decade AI has started to move from research into industry, and

quickly use of AI algorithms has spread. With the spread of AI systems into more work fields and entertainment platforms issues as the ones discussed above have arisen. Adoption of AI systems is still rising, more and more people are coming into contact with AI systems in more and more facets of their everyday life. With greater exposure comes greater opportunity for harm to occur.

Also, the development of new AI applications and technologies, despite their many benefits, still bring risks with them. This can already be seen in the case of chatGPT and it's sudden pervasiveness. It has not been released for a year yet, and already it is having far reaching implications for society. It has been known to display bias and false information at times, which could negatively affect humans and their dignity [18, 19].

One contribution for worry of harm to human dignity in the future of AI is the autonomy-first thinking. The belief that with enough data anything can be solved and humans can be replaced is still fairly common within AI. Pushing AI forward with this mentality is bound to perpetuate the issues of AI that are currently visible with bias, mistakes and danger to human autonomy [4].

Currently, regulation on AI is also behind and does not adequately protect human dignity when it comes to autonomous technologies. There are but few countries that have some regulations regarding AI in place [1].

Time for Change?

Following from the previous two premises, my third premise can be formed. Namely, something needs to be done if we want the field of AI to advance in a safe manner that preserves dignity. AI applications are sociotechnical systems. They are intertwined with the fabric of society now and will be even more so in the future. They mediate relations between humans and the world around them, both in the real and digital world [20]. As such their values should be aligned to ours if we want the benefits to outweigh the harms. In hurting human dignity AI systems can have a negative impact on our society, which we of course do not want. Because of this the European Group on Ethics in Science and New Technologies has even stated that the values on which our digital society is structured should be renewed with human dignity at the center [21].

Furthermore, the loss of human control over certain fields or decisions can have negative impacts on safety, responsibility, consent, institutional stability and of course human dignity [22]. This would be the course we are heading if the automation-first view stays dominant. A new way of conceptualising the role of AI in society is needed, together with guiding principles for design and governance which will ensure that human dignity is protected [4, 23].

The Human Centered Approach

HCAI for Human Dignity

The final premise leading to the main argument is that HCAI provides the right framework to guide the field of AI towards protecting human dignity. The goal of the HCAI framework is to put human needs and values at the center in the design, research and innovation of AI. HCAI systems are designed in such a way that they amplify human abilities and self-

efficacy instead of replacing humans with automated systems. One of the central concepts in the HCAI framework is the view that both high automation and high human control are possible and that either excessive machine or human control should be avoided. Through designing systems that allow for user control where appropriate the autonomy of users can be preserved [3,4]. Specifically this guards the self-worth as respect for autonomy conception of human dignity. In the HCAI framework human centered user interface design is paramount in achieving the right balance of human control and automation. It follows the Human-Control Mantra: a preview first, the human selects the action, action is initiated and the execution can be viewed. An example of this is digital cameras, here despite automation of focus and brightness the user has a lot of control. Here the user gets a preview before executing (taking the photo). They also can choose to change the focus if necessary or apply other settings. After execution the product can be reviewed [24]. Applying such thinking to AI systems, where intuitive interfaces are applied to support human autonomy, will help make AI systems more trustworthy as the state of an autonomous system can be made more transparent.

By also thinking in the design where the automation is placed, human dignity can be preserved. For example, currently people in need of help, such as elderly or disabled people, are seen as objects to be managed or passive targets of automation thereby instrumentalising them. However, AI could instead be used to better manage which caregiver is connected to which person, help plan routes and schedules helping the social connections [4]. This thought through placement of automation can help guard the non-instrumentalisation conception of human dignity.

Next to this HCAI also focuses on system design that promotes fairness, accountability, transparency and explainability [4,25]. Incorporating this will help make systems less biased and empower their users by allowing them to understand what is happening. This can help with human autonomy, make systems safer so less harm is done to humans and help guard the protection of vulnerable classes of persons conception of human dignity.

HCAI also proposes a three level governance structure for the development of more reliable systems, promoting safety culture in organisations and industry wide certification for trustworthy AI systems. In order to implement this collaboration from AI researchers, software engineers, business managers, government regulators, independent oversight committees and users of the products is needed. In doing this better designs can be researched by fostering the contact between users, industry and research. Regulations that help protect human dignity can be put in place and independent oversight can help enforce this [25,26]. By assuring the alignment with human values and by creating safe, reliable and trustworthy systems human dignity as protection of human species is better protected as well.

HCAI provides a framework for designing, implementing and regulating AI technologies such that human values and dignity are respected whilst fairness and autonomy are promoted. With HCAI pathways to these goals can be further researched and realised [3,4,25].

Going Human Centered

So is it necessary to adopt a human centred approach to AI? I have shown that given the current and future risk of AI to human dignity something does need to change. I have also elaborated on how HCAI can provide a framework that helps protect human dignity. Thus

in order for the field of AI to advance in a safe manner a human centered approach should indeed be adopted, because it protects human dignity.

A full discussion on how a human centered approach can be best adopted is outside the scope of this article. However, what stands out is that in order to move towards a human centered approach to AI adoption and collaboration of everyone involved in the AI life-cycle is needed. In research more attention should be given to researching fruitful ways of realising the HCAI vision further. The field of human computer interaction already provides some basis which can be applied to AI and further developed. Deeper collaboration of other research fields with AI is also necessary, such as psychology, cognitive science and social science. Research on HCAI is starting to grow in academia, however more attention is still needed. In industry adopting HCAI would mean making design more thoughtful about human values and needs. Especially moving away from pure machine control and allowing more control by users where appropriate. In [26] it is fully discussed what kinds of changes in management and company culture are needed to realise the HCAI view and produce safe, reliable and trustworty systems. In this paper the role of governance and independent oversight to ensure trustworthiness are also discussed. Lastly, there is also a role for the users of AI systems to demand their systems align with their values and their human dignity is not attacked. When HCAI is to be adopted, they also have a role in providing input to help better the systems.

Adoption of HCAI would entail creating a future where AI systems support humans, and humans are in control. Human dignity can be preserved and possibly even enhanced with this application of AI systems. The framework that HCAI provides is open to addition and can be extended in case new threats to human dignity are found [3, 4].

Objections and Replies

Now I will consider and reply to some possible objections to HCAI and the view I just presented.

One might object that by focusing on how to center AI on humans, progress of making AI more intelligent will be stunted. It could be argued that HCAI would limit the field of AI and stunt innovation as the focus on human values and dignity could prevent new and transformative AI technologies from being developed and applied. A case could be made that limiting the field like this, great benefits of such technologies will be missed out on if we adopt a human centered approach to AI. To this I reply that by adopting a human centered approach progress will not be stunted, it will even be accelerated. By adopting the design and business practices of HCAI constant innovation and betterment is achieved as constant monitoring and collaboration with research and users is implemented. The innovation taking place here will be in a direction that fits with our human values. The innovation might not be as rapid as it is now. However, this will ensure that the new technologies are safe for us to use and will not have worse consequences in the future [3, 4, 27].

Next, one might object that the adoption of HCAI specifically might not be necessary. As companies have put out principles and ethics guidelines for the development of AI. In these documents it is often mentioned that autonomy, non-maleficence and explicability are important. One might argue that with just this guidance, companies will solve these problems themselves. However, these guidelines currently fail to generate actual change. They are

currently used more as a way of ethics shirking by businesses [28]. HCAI would require a more holistic approach where everyone is taken on board with changing and working towards implementing AI that guards what is put forward in these principles and ethics guidelines. It will help actually putting theses guidelines into practice and thus is necessary in order to actually move towards AI that protects human dignity.

And lastly, in the HCAI framework speaks about centering technology on humans, to empower and protect them. However, who exactly is this universal human, how is this conceptualised? Who decides what is good for humans in general? It can quickly become problematic if this universal human is taken to be just westerners or approached from a limited point of view. Effectively only guarding the human dignity of a select group. HCAI does try to avoid this, in [4] and other works on HCAI the importance of including multiple perspectives is stressed. Taking into account cultural differences in designing AI systems is going to be challenging. Though HCAI does aim to achieve this, with inclusion of culturally diverse user bases and collaboration of different scientific fields to incorporate diverse cultural and ethical perspectives into AI.

Conclusion

In this paper I have shown the need to adopt a human centered approach to AI. This was done by first arguing that currently AI is hurting human dignity for all conceptions discussed in this paper. I have consequently argued that the risks to human dignity from AI will only increase in the future as AI technologies permeate deeper into the fabric of our society. Thus following that something needs to be done in AI research and application to stop AI from harming human dignity. I have also argued that HCAI provides the right framework to tackle these issues in AI. That it can provide us with the way to protect human dignity. Thus leading to the main argument of needing HCAI to lead us to a future where AI is not harming us and we can enjoy the full potential of AI systems. Now is the time to shape what the future role of AI will be in our lives and in our societies. So let us make it one that is dignified.

The Author

Marin Beijerbacht is currently a masters student in the Artificial Intelligence program at Utrecht University. This paper has been written for the course philosophy of AI and in cooperation with the Wish Will Way foundation. If you want to know more about the foundation visit https://wishwillway.org/. To connect with the author you can send a message on LinkedIn [1].

References

[1] M. L. Littman, I. Ajunwa, G. Berger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, *et al.*, "Gathering strength, gathering storms:

[1] https://www.linkedin.com/in/marin-beijerbacht-741254184

The one hundred year study on artificial intelligence (ai100) 2021 study panel report," *arXiv preprint arXiv:2210.15767*, 2022.

[2] N. Al-Rodhan, "Artificial intelligence: Implications for human dignity and governance," *Political Review*, vol. 27, p. 2021, 2021.

[3] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[4] B. Shneiderman, *Human-centered AI*. Oxford University Press, 2022.

[5] I. Kant, *Foundations of the Metaphysics of Morals*. Bobbs-Merrill, subsidiary of Sams, 1965.

[6] R. Debes, "Dignity," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Spring 2023 ed., 2023.

[7] R. Macklin, "Dignity is a useless concept," 2003.

[8] P.-A. Rodriguez, "Human dignity as an essentially contested concept," *Cambridge Review of International Affairs*, vol. 28, no. 4, pp. 743–756, 2015.

[9] W. B. Gallie, "Essentially contested concepts," in *Proceedings of the Aristotelian society*, vol. 56, pp. 167–198, JSTOR, 1955.

[10] C. McCrudden, "Human dignity and judicial interpretation of human rights," *european Journal of international Law*, vol. 19, no. 4, pp. 655–724, 2008.

[11] S. A. Teo, "Human dignity and ai: mapping the contours and utility of human dignity in addressing challenges presented by ai," *Law, Innovation and Technology*, pp. 1–39, 2023.

[12] P. Rey and W. E. Boesel, "The web, digital prostheses, and augmented subjectivity," *Routledge handbook of science, technology and society*, pp. 173–188, 2014.

[13] R. Srinivasan and A. Chander, "Biases in ai systems," *Communications of the ACM*, vol. 64, no. 8, pp. 44–49, 2021.

[14] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[15] D. Susser, B. Roessler, and H. Nissenbaum, "Online manipulation: Hidden influences in a digital world," *Geo. L. Tech. Rev.*, vol. 4, p. 1, 2019.

[16] D. Davidson, "Facebook targets 'insecure' young people to sell ads." https://www.theaustralian.com.au/business/media/facebook-targets-insecure-young-people-to-sell-ads/news-story/a89949ad016eee7d7a61c3c30c909fa6, 2017.

[17] D. Lim, "Killer robots and human dignity," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 171–176, 2019.

[18] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," *arXiv preprint arXiv:2304.09103*, 2023.

[19] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.

[20] R. Calo, "The scale and the reactor," *Available at SSRN*, 2022.

[21] E. Commission, D.-G. for Research, Innovation, E. G. on Ethics in Science, and N. Technologies, *Statement on artificial intelligence, robotics and 'autonomous' systems : Brussels, 9 March 2018*. Publications Office, 2018.

[22] J. Davidovic, "On the purpose of meaningful human control of ai," *Available at SSRN 4108171*, 2022.

[23] P. Inverardi, "The challenge of human dignity in the era of autonomous systems," *Perspectives on Digital Humanism*, pp. 25–29, 2022.

[24] B. Shneiderman, "Human-centered ai: A new synthesis," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part I 18*, pp. 3–8, Springer, 2021.

[25] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109–124, 2020.

[26] B. Shneiderman, "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[27] D. Tjondronegoro, E. Yuwono, B. Richards, D. Green, and S. Hatakka, "Responsible ai implementation: A human-centered framework for accelerating the innovation process," *arXiv preprint arXiv:2209.07076*, 2022.

[28] J. Qadir, M. Q. Islam, and A. Al-Fuqaha, "Toward accountable human-centered ai: rationale and promising directions," *Journal of Information, Communication and Ethics in Society*, 2022.