Aug 11th, 12:00 AM

Identifying and addressing unintended values when designing (with) Artificial Intelligence

Niya Stoimenova
Delft University of Technology, The Netherlands

Maaike Kleinsmann
Delft University of Technology, The Netherlands

Citation

Stoimenova, N., and Kleinsmann, M. (2020) Identifying and addressing unintended values when designing (with) Artificial Intelligence, in Boess, S., Cheung, M. and Cain, R. (eds.), *Synergy - DRS International Conference 2020*, 11-14 August, Held online. https://doi.org/10.21606/drs.2020.222

DRS2020
BRISBANE, 11–14 AUG
SYNERGY

Identifying and addressing unintended values when designing (with) Artificial Intelligence

Niya STOIMENOVA[a*], Maaike KLEINSMANN[a]

[a] Delft University of Technology, The Netherlands
[*] Corresponding author e-mail: n.stoimenova@tudelft.nl
doi: https://doi.org/10.21606/drs.2020.222

Abstract: A fundamental shift in the way society operates is approaching due to the prevalent adoption of self-learning technology like Artificial Intelligence (AI). Defined by the well-pronounced incongruence between their initial purpose and the values delivered to their multifaceted users, AI-powered systems are already being deployed in crucial social institutions such as hospitals, banks and courtrooms. To solve this tension, we first identify design practices that are suitable for the context of AI. Then we introduce a framework of three logical inferences that could aid designers to deliberately and continuously identify and address unintended values AI-powered solutions produce. The paper is concluded by three directions for future research.

Keywords: artificial intelligence; unintended values; purpose

1. Introduction

In the past few years we have witnessed impressive achievements of artificial intelligence (AI) [1] in variety domains such as speech recognition, visual object recognition, object detection, drug discovery, physics and genomics (LeCun et al., 2015; Ching et al., 2018). However, multiple cases exist in which an AI-powered solution perpetuated biases and behaved in unintended, and possibly unanticipated[2], ways (Caliskan et al., 2017). For instance, in August 2019, Apple officially released a new type of credit card that "represents all things Apple stands for. Like simplicity, transparency and privacy." (Apple, 2019). In

1 While the whole field of AI is particularly interesting and there are multiple speculations about the level of consciousness an artificial agent can attain (see e.g. Shulman and Bostrom (2012)), in this article we focus on the type of AI that is currently making the biggest strides and affecting people's lives the most – machine learning (ML). However, in the remainder of this article, we liberally use the terms model and algorithm, as well as AI and ML interchangeably for ease of explanation.

2 The word unanticipated implies that a behaviour was not foreseen and unintended is used to mean that while the consequence might have been anticipated, it was never intended. For the sake of simplicity, however, in this paper we consider unintended and unanticipated consequences to be the same.

November, however, a series of Twitter posts detailing a flaw with the AI behind the card became viral. It all started with a tweet from David Hansson (co-founder of Basecamp) who claimed that the algorithm gave him 20 times the credit limit his wife was afforded, despite the fact that she has a better credit score than him (Natarajan & Nasiripour, 2019). Even Apple's co-founder, Steve Wozniak, tweeted the following:

> The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019. — Steve Wozniak (@stevewoz) November 10, 2019

Yet, such AI-powered systems are already being deployed in crucial social institutions such as hospitals, banks and courtrooms (Crawford & Calo, 2016). Due to their extreme diversity, ubiquity and complexity, however, anticipating their behaviour remains challenging (Rahwan et al., 2019). As a result, a wide array of practitioners and scholars have been warning against the broad, unintended consequences of AI-powered solutions that can shape human behaviours and societal outcomes in both intended and unintended ways (Rahwan et al., 2019).

The aim of this paper is neither to discuss potential dystopian futures nor to convince the design community of ways AI could improve human health, safety, productivity or even creativity. Rather, its aim is to identify and elaborate upon possible ways for design to address what we consider to be the major tension of AI-powered solutions interacting with their environments – the incongruence between the purpose (what is intended/anticipated) of the solution and the value[3] it delivers which might not always be "aligned with human interests" (Soares & Fallenstein, 2014; Taylor, 2016). For instance, while the purpose of an Apple Watch is to track your heartbeat, the values it delivers to each stakeholder may vary – from saving someone's life to changing how heart disease is diagnosed and thus affecting the market for diagnostic equipment, changing the cardiologists' job or even failing to alert for heart anomaly in a context where people blindly trust it and consequently contribute to someone's death.

As such, this paper is structured as follows, first we unpack the current state of design's involvement in ensuring the delivered values are aligned with humans' interest. Then we identify additional design practices that could help us to address the tension. These practices are combined in an initial framework of logical inferences on how to design for identifying and addressing potential unintended values early in the lifecycle of AI-powered solutions. The paper is concluded by an elaboration upon three viable directions for future research.

2. Aligning values with human interests through design

One of the most prominent emerging ways to ensure values created by AI-powered solutions align with humans' interests is the so-called human-centered machine learning (HCML).

3 We consider value to have a broader meaning that includes dimensions such as practical, economic, ethical and aesthetic rather than only moral values or (economic) utility. As such, the value we describe is aligned with the definition of value-in-use elaborated upon by Vargo and colleagues (2008).

HCML highlights the need to take a human-centered perspective on how ML solutions impact people (Ramos et al., 2019). The approach is based on principles, guidelines and strategies for designing user interactions with AI-powered solutions developed in the Human Computer Interaction (HCI) community in the past twenty years (Amershi et al., 2019). As such, to our knowledge, its efforts are geared towards making the output of an AI model easier to understand by its users, ultimately ensuring a seamless user experience and personalisation (PAIR, 2019). Currently, companies like Google, IDEO and Microsoft are spearheading the efforts in this area with recent releases of Google's People + AI handbook, IDEO's set of AI ethics cards (Sampson & Chapman, 2019) and Microsoft's design principles (Amershi et al., 2019). Academia is slowly catching up. In November 2019, the number of articles appearing on Google Scholar that have the following strings in their titles: "human-centered AI", "human-centered ML", "human-centered machine learning", "human-centered artificial intelligence", amounts to almost 500 with scholars from the HCI field thus far taking the lead.

As already mentioned, however, one of the biggest problems with current AI-powered systems is the fact that they create values that are not intended and can ultimately inflict harmful impact on humans and humanity. This problem is not specific to AI. In fact, technology and design have a long-standing tradition of producing unintended values (i.e. the typewriter, microwaves, the Internet, Facebook's like button). According to Ihde (2018), the notion that one can design into a technology its purposes and uses is a fallacy since the designer's intent may be subverted, become a minor use, or produce completely unintended results (Ihde, 2008). Even human-centered design practices have been repeatedly shown to lead to unintended consequences or the so-called "dark patterns" of design (i.e. the "pull to refresh feature") (Gray et al., 2018). As AI-powered agents gain autonomy and act in more complex domains, it may become progressively harder to anticipate the impact and implications of the array of possible unintended values (Amodei et al., 2016). Therefore, before carrying over existing design practices and methods to designing (with) AI, attention should be given to whether the limitations such exhibit could get amplified when transferred to the new context. Moreover, to be able to create a principled approach to identifying and addressing unintended values early in the solution's lifecycle, we should critically asses existing practices and identify the ones that could help us to deal with the purpose-values tension.

We contend that the criteria for such design practices are three: (1) they should address the incongruence between purpose and (unintended) values, (2) they should proactively anticipate potential future values and (3) they should make a clear distinction between the way a system behaves and the way it is being used. The criteria are interconnected and continuously amplify each other. For instance, an AI-powered solution is created with a certain purpose (e.g. to monitor the heartrate of its user). However, the way it is used influences what it learns and consequently its behaviour might be novel. This makes the future values it can deliver difficult to anticipate and analyse (Rahwan et al., 2019).

Nota bene. *A clear distinction between biases and values has to be established, however. The majority of AI models are trained on implicitly biased human-generated data. Moreover, they are designed to be biased towards the variables that will allow them to achieve their goal (Dixon et al., 2018). To exemplify that, let us speculate about the Apple Card example mentioned above. The model trained to define credit limit is instructed to be biased towards the probability of people paying their credit card bills on time. As such, it parses through a large database of past credit card transactions. The problem arises, therefore, once the model that is not explicitly instructed to discriminate between the gender or socio-economic background of its users, starts doing so. This is called unintended bias. Unintended values are produced when this unintended bias starts to affect the use of the system and in particular when different individuals are treated differently (Dixon et al., 2018). Therefore, while biases are inherent to the data and the design of the model, it is their manifestation, once they start to interact with their context, that produces unintended values. As such, we deliberately choose to focus on the values since latent biases are notoriously difficult to identify when dealing with non-binary data (Bellamy et al., 2018). Moreover, we do not wish to label all biases as bad since such cognitive heuristics play an important role in human cognition. Therefore, instead of directly trying to identify biases, we focus our efforts on identifying their observable results: the created values.*

A prominent theory addressing the notion of purpose, the way a solution behaves and the way it is being used is that of *innovative abduction* elaborated upon by Roozenburg in 1993. According to him, the design process always starts with a purpose. For instance, when designing a kettle, the purpose would be to "be able to boil the poured-in water". To fulfil its purpose, however, the kettle needs to behave in a certain way (e.g. the bottom needs to heat up, so it can transform the heat to the water inside). Roozenburg terms this **mode of action** and uses it to signify the (functional) behaviour of the artefact in response to influences exerted on it from its environment. It is our contention that this notion is well-adjusted to represent the way an AI-powered solution behaves, since, according to Roozenburg, the mode serves as the bridge between the artefact and its immediate environment. To account for the user's action on the product (e.g. putting the kettle on the burner), Roozenburg introduces the notion of **actuation**. Such signifies the continuous action of a user that allows the artefact to function and be 'connected' to its immediate environment. Finally, the **form** of the artefact should also be considered. Therefore, the form of the kettle and the way it is used (actuated) causes it to behave in a certain way (mode of action), and therefore, by this behaviour, it can fulfil its purpose. The logical expression of the innovative abduction process is, therefore, as follows:

((form and actuation) → mode of action) → purpose

As such, Roozenburg's theory accounts for **purpose**, the way a solution behaves **(mode of action)** and the way a solution is being used **(actuation)** and elaborates upon the relationship amongst them. However, he does not deliberate on how value is created. To search for guidance in this direction, we consider the theory on abductive reasoning Dorst (2011) introduced. Partially iterating on Roozenburg's work and reflecting the changes in the field,

according to Dorst (2011), value is the result of the following inference:

WHAT + HOW → VALUE

The WHAT in the formula represents an object, a service, a system – a wider domain of application than Roozenburg's Form. The HOW signifies the working principle of the solution. The sum of both leads to (an aspired) VALUE that is known at the start of the design process and therefore, intended. To ensure that this intended Value results from the combination of the What and the How, Dorst introduces the frame creation practice (Dorst, 2015). He defines it as a cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved. Therefore, it delivers the intended value the designer strives for (Dorst, 2011).

These two theories allow us to identify the design practices that could address the elements of value, purpose, mode of action and actuation[4]. However, neither of them fulfils our second criterium: *"they should proactively anticipate potential future values"*. A design practice that could allow us to anticipate future values and monitor changes over time is prototyping. Prototypes are widely recognised as an important means to explore and communicate what it will be like to interact with future products, systems and services (Buxton, 2007; Lim, et al., 2008). They are oftentimes used to help designers learn, discover, generate, and refine designs (Buxton, 2007) by stimulating framing, and discovering possibilities in a design space (Lim et al., 2008). As such, they can play different roles– evoke a focused discussion in a team, test hypotheses, confront theories, allow users to experience their world differently (Sanders & Stappers, 2014) or generate deep level of understanding about novel contexts (Lim et a., 2008). Therefore, they can minimize design errors that may otherwise occur late in the process (Deininger et al., 2017). While many definitions of a "prototype" exist, depending on the design field (i.e. Sanders & Stappers, 2004; Lim et al., 2008; Pei et al., 2011), instead of defining what a prototype could look like, we adopt the notion that prototypes should provide help in discovering new aspects of the problem at hand and support the invention of design requirements (Schön & Wiggins, 1992; Suwa et al., 2000). As such they can take any form, shape, and appearance, based on the choice of material (Lim at al., 2008; Deininger et al., 2017) and the phase of the design process in which they are being used.

3. Initial Framework

To present and explain the initial framework, we will use a fictional example of a designer who wants to create solutions that reduce the burden people with chronic kidney problems

4 We referred only to these core papers of Roozenburg (1993) and Dorst (2011) because they are very explicit in their definition of abductive reasoning in the field of design while maintaining a broader orientation (they adopted theories from e.g., March (1976), Habermas (2015) and Schön (1983)). The most important conclusions we draw here have been checked with other (later) papers by Roozenburg, Dorst and other authors (i.e. Kroll and Koskela (2017)). Moreover, the formulation of our framework was also informed by the work of Dong (i.e. Dong and MacDonald (2016), Takeda (i.e. 1994; 2001) and the work on Function-Behavior-Structure ontology of Gero (i.e. 2007). Last but not least, we took a conscious decision to only include these two theories for the sake of readability and theoretical coherence.

feel[5]. Therefore, our purpose becomes *"to ensure people with chronic kidney conditions are in control of their health".* That would mean that a process of transformation needs to happen from the current situation of kidney patients spending a considerable amount of time undergoing dialysis, lacking an overview on their daily progress and not knowing whether the problems they experience are life-threatening or simply mild ones, to the preferred one of being informed and in control of their health.

To be able to design a solution that fulfils the intended **purpose**, the first step will be to collect **data** (both qualitative and quantitative). This process could go as follows: we would carefully study our patient's context through regular interviews and observations, and then map their day-to-day journey. These will provide contextual understanding of the problem and help us to discover, for instance, that physicians are oftentimes unsure about the precise dosage of medications since each one of them can directly influence the blood potassium levels of the patient. Higher levels than recommended can be fatal and require immediate treatment. Yet, the only way to identify blood potassium levels is for the patient to undergo an invasive test performed in a laboratory. This produces high levels of uncertainty and stress for the patient and all stakeholders involved.

We, then, build upon this insight with the medical research knowledge that potassium levels can be detected in an electrocardiogram (ECG) (Topol, 2019)). Subsequently, we should collect a database of ECGs. This combination of **purpose** and **data** allows us to define an interesting vantage point (a frame) from which we can achieve our purpose or:

purpose + data → frame

Going back to our example, a possible frame in our situation could be that "*if potassium levels are detected regularly, changes in the dosage of a medicine can be easily administered in the comfort of the user's home (no need for blood tests)*".

The next step is to explore different modes of action our solution can exhibit within the frame we created. We define the mode of action as the way an AI-powered solution attempts to influence its environment. Such mode is always associated with the identified frame. Therefore, for our example, a plausible **initial mode of action** is: "*deep neural network detects potassium levels in ECGs"*. The combination of the mode and the frame will help us to understand how a prototype can be manifested. We specifically refer to the result of this inference as a prototype instead of a "What" (Dorst, 2011) or a "Form" (Roozenburg, 1993) since, as discussed, prototypes allow us to anticipate future values. Moreover, since AI-

5 Although this is a fictional example, to provide the necessary level of detail, we use the case of the company AliveCor. The start-up produces cell-phone cases and AppleWatch wristbands that can perform electrocardiograms (ECGs) (Topol, 2019). Based on them, potassium levels in "near real time" can be detected without drawing blood (Dillon & Friedman, 2018). This case was deliberately chosen as it addresses the complex context of healthcare in which multiple stakeholders (e.g., patient, nephrologist, GP, hospitals, hospital staff, medical device systems manufacturers, insurance companies) come into play. They all expect the solution to deliver values tailored to them, while the AI-powered solution continuously learns from its users and consequently could exhibit novel behaviours. Due to the complexity of the context, potential values are difficult to anticipate. As such, this case provides a wide range of challenges and exemplifies the type of contexts and domains for which our framework is developed.

powered solutions continuously learn and adapt to their contexts, we contend that viewing each of their states as a prototype will promote the notion of designing for something transient (i.e. a solution that can always change the values it delivers). In effect, this will enable the designer to learn continuously as the solutions evolves. This leads us to:

frame + mode of action → prototype

A prototype resulting from this inference might be manifested as a simple piece of hardware (e.g. a strip) equipped with electrodes that can measure pulse. It is important to note that at this initial stage, the prototype we make use of is generative[6] as we still do not have a clear hypothesis neither on how the strip will be used, nor on the values it will produce. Therefore, we need a prototype that will aid us to generate such hypotheses. Since the goal of this prototype is to better understand the context, a good starting point would be to provide the users with the strip and without much guidance to observe the way they use it over the course of a month. Doing so will generate multiple insights on the context, time of day, actuation and expectations users have while wearing the prototype. As well as on the way it affects and impacts their daily routine and that of the already identified stakeholders. Therefore, it is through this first prototype that we understand the ways it can be actuated and ultimately observe the different values it can create. When put in a formulaic expression, this appears as:

prototype + actuation → values

The value here is not necessarily the same as the purpose and can be different for each user and stakeholder. In our case, it could be that we observe that the strip creates new dynamics in our patient's life by introducing a lot of uncertainty and tension between the patient and her partner as they begin to obsess over even insignificant changes registered by the strip. While this value stems from the fact that the patient is more in control of her life, it creates unintended and undesired values – i.e. tension between her and her partner.

This insight will trigger another loop through our framework. Therefore, we start collecting new data (both qualitative and quantitative) on how the patient's daily routine changed, interviews with the patient and her spouse, as well as with the involved stakeholders. But also, reviewing the ECG and potassium levels and map them to events of what was observed and communicated in the previous loop. This new iteration of data collection will help in refining our frame and add another dimension to the mode of action by adding behaviours that could address the identified tension. These are reflected in a new, more detailed, prototype. Consequently, gradually new intended and unintended values are uncovered in later iterations. As such, also the prototypes we design could be more detailed or help

6 Apart from using prototypes as a means to evaluate design's failure or success (Lim et al., 2008), they can also be used as a tool for learning, discovering, generating, and refining designs (Buxton, 2007) by stimulating framing, and discovering possibilities in a design space (Lim et al., 2008). Moreover, using externalisations in a generative manner is not new to the scientific practice. According to Magnani (2007), for instance, they can be used to capture the part of scientific thinking in which the role of action is central, and the features of this action are implicit and difficult to elicit. He terms this manipulative abduction.

us to understand a different part of the solution such as the way to deliver information to nephrologist, GPs and patients. Therefore, the effect these delivered values have not only on the patient, but also on the community and the other stakeholders (i.e. the GP, nurses, family) can be studied. Doing so will ensure that more details about the interaction between the unintended values and the context they operate in could be generated. Moreover, the collected insights might necessitate the creation of a different, more accurate frame and/or mode of action. In some cases, the purpose also could be refined or even reconsidered. Thus, we contend that designing (with) AI-powered solutions could go through a continuous loop of this framework:

purpose + data → frame

frame + mode of action → prototype

prototype + actuation → values

The deliberate choice whether and how the solution should be changed to address the identified values resides with the designer or the multidisciplinary team that designs the solution. While at different times, different variables might come into focus, this framework will allow the team to understand the context and proactively anticipate the possible implications of their design. In fact, it is our contention that this framework will allow the teams designing such solutions to identify and address potential unintended values early in the solution's lifecycle. Furthermore, once the solution is deployed, the variables dedicated to each aspect of an AI-powered solution will make the monitoring and eventual explanation and troubleshooting easier.

4. Future Research Directions

The framework we present is explorative and purely theoretical in nature. It is our contention, therefore, that empirical research on its implementation will help us better understand its nature and the manner in which it could identify and address the unintended values AI-powered solutions create. As such, we have identified and outlined below three possible directions for future research.

4.1 Identifying values

Being able to continuously identify and address values that the ever-adapting AI-powered solutions provide to their users is of paramount importance. It is our contention, therefore that an important distinction has to be made between two variables: V_i denoting the intended values and V_u which represents the unintended ones. For instance, the solution delivers a value that ensures that the person with kidney problems is in control of her life by giving her an overview of her daily activities and providing tips how she can improve. As such, it stays within the boundaries of the initial purpose and can be denoted as V_i. If the solution, however, suddenly starts to deliver values outside of the intended ones like the one we described earlier, the delivered value falls within V_u. These unintended values, however,

could be both positive and negative. Nevertheless, to be able to identify such values, we first need to carefully study their nature, the way they emerge and evolve, the manner in which they could be calculated, and the role designers should play in deliberately steering the solution towards certain values. While these should always align with human interest, a way to consider the different stakeholders should be designed and further studied as well.

4.2 Prototyping

Our framework also directly deals with the uncertain implications AI-powered solutions have on society by continuously re-examining the problem and solution space (Dorst & Cross, 2001) through the practice of prototyping. This makes the proposed approach different from the notion of HCML which is based on purely inductive and deductive ways to test hypotheses (i.e. Google's PAIR handbook, 2019). In our framework we propose to use prototyping not only as an evaluative (analysis) means, but also as a generative (synthesis) one. As already explained, using prototypes in a generative manner can create communicable accounts of new experiences that can be integrated into previously existing systems of experimental and linguistic (theoretical) practices.

The importance of using generative prototypes is best exemplified by the case of Microsoft's bot Tay who "emulates the casual, jokey speech patterns of a stereotypical millennial". The aim of the bot was to learn from her conversations with people on Twitter and get smarter over time. Even though the team worked with comedians and extensively tested the bot with users before being deployed, (Lee, 2016), it quickly started generating racist tweets, defending white supremacy, denying the Holocaust and praising Nazis (Price, 2016). Testing the bot inductively and deductively before its release did not allow its creators to envision all possible ways in which the bot could be "attacked" (Lee, 2016).

Our deliberate effort to explore the context through a loop of generative and evaluative prototypes allows —in theory— for better understanding of the possible unintended values. Therefore, further empirical research is necessitated to corroborate this statement and it is our contention that those should try to address the following challenges: 1) prototyping with AI is challenging since the performance of a deployed AI system can constantly fluctuate and diverge when it gains new data to improve its learning. Consequently, it seems that established prototyping methods cannot address this new context (Yang et al., 2020); and 2) prototypes are usually intended for and tested with individuals. Yet, the majority of recorded unintended values of AI-powered solutions happen when the individual interacts with her context and community.

4.3 New methodologies

Building on our previous point, it is our contention that a fruitful third area for further research could be the investigation of new research methodologies that allow us to combine the language-based research methods for studying design activities such as protocol analysis (see e.g. Cross, 2001; Hay et al., 2017) and/or natural language processing methods (see

e.g. Dong, 2005) with such that allow design researchers to incorporate non-verbal design activities such as prototyping. Doing so would not only generate insights about the possible approaches to design (with) AI, but also about the different processes under the umbrella of design (Kleinsmann & Ten Bhömer, forthcoming). A steppingstone in this research direction could come from the work of Cramer-Petersen and colleagues (2019) who analysed the reasoning patterns in design at a micro-level. According to them, doing so holds the potential to advance the understanding of design activity and can be applied to develop support tools and methods given future research (Cramer-Petersen et al., 2019).

5. Conclusion

The aim of this paper was to identify design practices that could address the incongruence between the purpose of an AI-powered solution and the (unintended) values it delivers to its multifaceted users. To address this tension, we identified three criteria that design practices and methods should fulfil if they are to be used in the context of AI: (1) address the tension between purpose and (unintended) values, (2) proactively anticipate potential future values, and (3) make a distinction between the way a system behaves and the way it is being used. Using these as a guideline, we identified the elements of purpose, mode of action and actuation (Roozenburg, 1993), value and frames (Dorst, 2011) and prototyping (both to generate hypotheses and to evaluate them). We then elaborated upon the connection between purpose and intended and unintended values through a series of logical inferences (a framework). Last but not least, we introduced three directions for future research.

The tension between intended purpose and delivered values is not new to the design discipline. Nevertheless, it is our contention that it gains fundamental importance in the context of AI-powered solutions and should be deliberately addressed. When the discipline was primarily concerned with the design of a product, a clear line could be drawn to where the responsibility of a designer ends. If a designer creates a kettle, for example, and someone is murdered with it, the blame can clearly be assigned to the user who uses the product in an unintended way. However, in the case of Apple Card, despite using the service as intended by the designer, the user gets punished (i.e. a woman receives much lower credit limit). The responsibility and the blame cannot be easily assigned and, in most cases, the reason for the undesirable outcome cannot be readily understood either (Amodei et al., 2016; Rahwan et al., 2019). The important question is then: *"How are we, as designers, going to deal with this conundrum – do we hide behind the notion of technology we don't understand, or do we take full responsibility for all the unintended values we create?".*

6. References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019, April). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (p. 3). ACM.

Apple. (2019, August). *Apple Card*. Retrieved from Apple: https://www.apple.com/apple-card/

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Buxton, B. (2010). *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Xie, W. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15*(141), 20170387.

Cramer-Petersen, C. L., Christensen, B. T., & Ahmed-Kristensen, S. (2019). Empirically analysing design reasoning patterns: Abductive-deductive reasoning patterns dominate design idea generation. *Design Studies*, *60*, 39-70.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature News*, *538*(7625), 311.

Cross, N. (2001). Design cognition: Results from protocol and other empirical studies of design activity. In *Design knowing and learning: Cognition in design education* (pp. 79-103). Elsevier Science.

Deininger, M., Daly, S. R., Sienko, K. H., & Lee, J. C. (2017). Novice designers' use of prototypes in engineering design. *Design studies*, *51*, 25-65.

Dillon, J., & Friedman, P. (2018, March 26). *Using AI to Invent New Medical Tests*. Retrieved from Harvard Business Review: https://hbr.org/2018/03/using-ai-to-invent-new-medical-tests

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, 26(5), 445-461.

Dong, A., & MacDonald, E. (2016). From observations to insights: the hilly road to value creation. In *Analysing design thinking: Studies of cross-cultural co-creation* (pp. 465-481). CRC Press.

Dorst, K. (2011). The core of 'design thinking'and its application. *Design studies*, *32*(6), 521-532.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Gero, J. S., & Kannengiesser, U. (2007). A function–behavior–structure ontology of processes. *Ai Edam*, *21*(4), 379-391.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018, April). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 534). ACM.

Habermas, J. (2015). *Knowledge and human interests*. John Wiley & Sons.

Hay, L., McTeague, C., Duffy, A. H., Pidgeon, L. M., Vuletic, T., & Grealy, M. (2017). A systematic review of protocol studies on conceptual design cognition. In *Design Computing and Cognition'16* (pp. 135-153). Springer, Cham.

Ihde, D. (2008). The designer fallacy and technological imagination. In *Philosophy and design: from Engineering to Architecture*, by Pieter E. Vermaas, Peter Kroes, Andrew Light and Steven E. Moore, 51-61.

Kleinsmann, M., & Ten Bhömer, M. (forthcoming). The (new) roles of prototypes during the co-development of digital Product Service Systems. *International Journal of Design*, *xx*), xx-xx.

Kroll, E., & Koskela, L. (2017). Studying design abduction in the context of novelty. *The Design Society*.

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

Lee, P. (2016, March 25). *Learning from Tay's introduction*. Retrieved from Official Microsoft Blog: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lim, Y. K., Stolterman, E., & Tenenberg, J. (2008). The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(2), 7.

Magnani, L. (2007). Abduction and chance discovery in science. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *11*(5), 273-279.

March, L. (1976). The logic of design and the question of value. The architecture of form.

Natarajan, S., & Nasiripour, S. (2019, November 9). *Viral Tweet About Apple Card Leads to Goldman Sachs Probe*. Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2019-11-09/viral-tweet-about-apple-card-leads-to-probe-into-goldman-sachs

PAIR. (2019, May). *People + AI Guidebook*. Retrieved from PAIR: https://pair.withgoogle.com/

Pei, E., Campbell, I., & Evans, M. (2011). A taxonomic classification of visual design representations used by industrial designers and engineering designers. *The Design Journal*, *14*(1), 64-91.

Price, R. (2016, March 24). *Microsoft is deleting its AI chatbot's incredibly racist tweets*. Retrieved from Business Insider: https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?international=true&r=US&IR=T

Ramos, G., Suh, J., Ghorashi, S., Meek, C., Banks, R., Amershi, S., ... & Bansal, G. (2019, April). Emerging Perspectives in Human-Centered Machine Learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (p. W11). ACM.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Jennings, N. R. (2019). Machine behaviour. *Nature,* 568(7753), 477.

Roozenburg, N. F. (1993). On the pattern of reasoning in innovative design. *Design Studies*, *14*(1), 4-18.

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5-18.

Sampson, O., & Chapman, M. (2019, May 9). *AI Needs an Ethical Compass. This Tool Can Help*. Retrieved from Ideo: https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help

Schön, D. A. (1983). *The reflective practitioner: how professionals think and act.* New York: Basic Books.

Schön, D. A., & Wiggins, G. (1992). Kinds of seeing and their functions in designing. *Design studies*, *13*(2), 135-156.

Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, *8*.

Stumpf, S. C., & McDonnell, J. T. (2002). Talking about team framing: using argumentation to analyse and support experiential learning in early design episodes. *Design studies*, 23(1), 5-23.

Suwa, M., Gero, J., & Purcell, T. (2000). Unexpected discoveries and S-invention of design requirements: important vehicles for a design process. *Design studies*, *21*(6), 539-567.

Takeda, H. (1994, January). Abduction for design. In *Formal design methods for CAD* (pp. 221-243).

Takeda, H., Yoshioka, M., & Tomiyama, T. (2001). Roles and formalization of abduction in synthesis. In *Proceedings of the Annual Conference of JSAI 15th Annual Conference, 2001* (pp. 30-30). The Japanese Society for Artificial Intelligence.

Taylor, J. (2016, March). Quantilizers: A safer alternative to maximizers for limited optimization. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Hachette UK.

Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On value and value co-creation: A service systems and service logic perspective. *European management journal,* 26(3), 145-152.

Yang, Q., Steinfeld, A., & Zimmerman, J. (2020) Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. *In Proceedings of the CHI Conference 2020*

About the Authors:

ir. Niya Stoimenova is a Doctoral Candidate at TU Delft. Her research is multidisciplinary and focuses on: (1) early identification of the unintended consequences AI-powered solutions create and (2) the design of adaptive organizational practices that allow such consequences to be addressed swiftly.

Professor Maaike Kleinsmann studies the role of design in digital transformation. She develops data-enabled design methods to equip designers to digitally transform healthcare. She leads Cardiolab; a multi-stakeholder effort that collaboratively reduces the burden of cardiac diseases through smart technologies.