Blind Spots in AI: the Role of Serendipity and Equity in Algorithm-Based Decision-Making

Cora van Leeuwen
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
cora.van.leeuwen@vub.be

Annelien Smets
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
annelien.smets@vub.be

An Jacobs
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
an.jacobs@vub.be

ABSTRACT

Decisions support systems (DSS) are used more and more to offer right information at the right time. Serendipity has been pro- posed to ensure that the experience is broad and engaging. However, only designing for serendipity might not be enough to avoid historical discrimination affecting your DSS. For this reason we argue to include equity when designing for serendipity.

1. INTRODUCTION

Managing knowledge is an important skill, it can reduce or enhance the power of individuals and organizations [44]. Nowadays, systems based on big data and algorithmic processing are increasingly applied to support knowledge management and resulting activities. For example, algorithmic matchmaking systems are being used to match job seekers and potential employers, and several approaches have been put forward to implement these complex tasks [3]. These decision support systems (DSS) are trained to offer the right information at the right time with high accuracy, making use of (big) data. However, we are increasingly becoming aware that there is a problem of bias and that it is a difficult problem to address. It emerges in the system before the data is collected as well as in the other stages of the deep learning processes [28]. The question of "how standards of unbiased attitudes and non-discriminatory practices can be met in (big) data analysis and algorithm-based decision-making" is consequently a timely one. Today, there are many tools focusing on technical solutions mitigating bias built up by choices in training data and modelling methods [39]. However, this only partially solves the challenge to enable the creation of both a performant fair and engaging, interesting system. After all, these tools focus on the known social or statistical biases. What with the unknown unknown, the blind spots? We agree with scholars like Reviglio [44] and Ge et al. [26] and argue that designing for serendipity could help to overcome these blind spots in DSS. However, we want to debunk the idea that designing for serendipity is a guarantee to develop a system free from any bias. We therefore argue that the implementation and operationalization of serendipity should take into account additional principles, such as equity. Hence, the particular question this paper aims to address, is why and how serendipity and equity can help to overcome blind spots in the design of algorithmic decision support systems (DSS). In the remainder of this paper, we will first elaborate on the notion of blind spots and discuss why we consider them to be a result of existing paradigms to reduce information overload. Second, we address the notion of serendipity to overcome these blind spots. Following this, we put forward three guiding principles to introduce equity in the design of DSS. Throughout this work, we demonstrate these principles by means of an hypothetical case example of a job-matching system that aims to help job seekers find interesting vacancies.

2. INFORMATION OVERLOAD

The increasing amount of available data and digital information systems provide great opportunities for these kinds of decision support systems. However, as discussed in related literature [29, 11], this also comes with the problem of information overload. Although existing recommender systems have shown to be an efficient remedy by applying personalization and filtering techniques, there are growing concerns about the potential drawbacks of these systems (e.g. [11]). Indeed, while the current paradigm has shown to be effective to reduce information overload, it is also being criticized to exploit convergent system behavior rather than cater divergent behavior [44]. This aligns with the emerging call from scholars for additional and/or alternative metrics to optimize these information systems beyond accuracy or relevance (e.g. [33]).

Perhaps the most contentious discussion in this regard is the one about filter bubbles in online (social) media. Here, the hypothesis is that algorithmic personalization focused on accuracy or relevance results in a diminished exposure diversity. The latter implies that users are only being exposed to information that confirms their beliefs or properly aligns with their preferences. In this way, users of the system are literally blind to any information outside their bubble. While this is an useful feature in some cases - a dog owner does not want to get recommendations for cat food - in other situations it might be detrimental. For example, when you look online for information about a particular topic, e.g. the usefulness of vaccines, you might only find information that confirms your existing beliefs and thus result in a confirmation bias [30].

While this is a striking example that many like to associate with notable events in our contemporary society, the problem of information filtering and algorithmic curation exceeds this single application domain. Indeed, as decision support systems are increasingly being used in healthcare, financial systems or juridical settings, these domains suffer from these biases too. Moreover, these blind spots are not only due to the filtering techniques themselves. They also result from the available information in the first place: the data sources and training data. In the case of job matching systems, for example, there is a recurrent observation that women are more often shown vacancies for part-time jobs, even though gender is not a variable that is taken into account in the model [51]. This appears to be, however, due to the actual situation in the job market in which women seem to more often apply to part-time jobs and consequently this behaviour is reinforced in the matchmaking model. The question then arises about how to deal with this: should the system simply keep this imbalance or should it try to remediate? And if we decide upon remediating, how can we make sure the system indeed includes items that might score less on relevance or accuracy, and thus represent the blind spots? The challenge is hence how to design these systems with attention to these blind spots? In the next section we illustrate how the concept of serendipity could be a first step in that direction.

3. SERENDIPITY AGAINST BLINDSPOTS

3.1 Serendipity in digital environments

The idea of having a set of guidelines that helps us to discover the unknown sounds appealing to many. Unexpected discoveries are a key driver for innovation and growth, and the study of how people encounter information accidentally has therefore been an important field in information science over the last decade. In this domain, the notion of serendipity is used to refer to the "unplanned ways to encounter resources that we find interesting" [10, p.7] . Serendipity is associated with a line of groundbreaking discoveries such as Alexander Fleming's discovery of Penicillin or Archimedes' principle. Indeed, this 'eureka moment' reflects the key characteristic of serendipity: an unexpected discovery.

Despite the importance of serendipity in epistemology, more recently the concept started to gain attention in digital information environments as well. Indeed, while the Web could be considered to be the ultimate serendipity engine [32], the problem of information overload and resulting algorithmic filtering techniques seem to diminish this serendipitous potential. Scholars therefore argue to apply serendipity as a key design principle for digital environments [44]. Serendipity is indeed concerned with discovering the unknown unknown and has been proposed to be applied in recommender systems to improve their quality and resulting user satisfaction [26]. Similarly, serendipity can play an important role in decision support systems to overcome the previously described blind spots. For example, adding serendipity to a job matching system, might result in recommending vacancies that the job seeker would never have thought of by him/herself. However, the question remains how to design these serendipitous encounters?

3.2 Designing for serendipity

While the notion of serendipity has gained attention in several scientific disciplines, ranging from psychology to sociology of science and computer science, there is a recurrent misconception that turns it into an ill-defined buzzword [45]. Indeed, serendipity is often referred to as a happy accident, an exceptional situation in which the right person was in the right place. However, it is important to understand that serendipitous discoveries are more complex: they are the result of a favourable combination of an individual's characteristics and so-called environmental affordances. The latter are "opportunities for action offered by the real world" [47, p.117] that allow individuals to engage in information-seeking activities leading up to serendipitous encounters.

3.3 The key affordances of serendipity

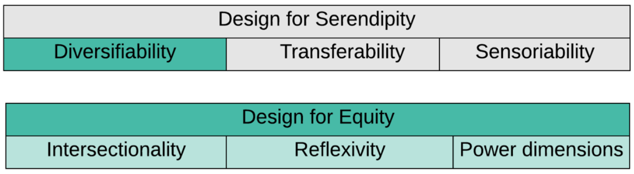
This approach allows one to overcome the paradox of 'designing accidents' as we will not design serendipity itself, but design for serendipity and thus a so-called serendipity potential. As mentioned before, there is an emerging line of research focusing on cultivating this serendipity potential in digital environments and stresses its importance and ethical value [43, 10, 44] . Björneborn [10] identifies three key affordances for serendipity: diversifiability, transferability and sensoriability. These affordances represent the "key aspects of human interactions with environments" [10, p.7]. The first one, diversifiability, refers to the extent to which such an environment can be diversified in terms of content. In the case of our job matching system, this would mean a set of recommended job vacancies across different sectors and/or with multiple modalities (both part-time and full-time for example). In this same example, the notion of transferability would refer to the capacity of the system to explore the possible vacancies. How easy can one navigate in between several vacancies; are there multiple ways to end up with one particular vacancy (e.g. through multiple search words), etc. For example, a vacancy for a 'store assistant' could be retrieved both when looking for jobs in retail as well as jobs with people. This illustrates one of the benefits of having divergent systems because they allow to create new semantic knowledge, in this case for the end-user. Finally, sensoriability deals with the sensory stimuli in the environment. In digital environments, this often refers to a combination of images, colored hyperlinks, explicit keywords, notifications or particular suggestions that are displayed.

3.4 Interaction of the system

An important additional aspect of these affordances is the identification of re- lated personal characteristics that are considered to be "the actoral components of these [environmental] affordances" [10, p.7]. More specifically, Björneborn [10] identifies three key personal factors that correspond to each of the previously described affordances: curiosity (diversifiability), mobility (traversability) and sensitivity (sensoriability). Without going into detail about each of these personal characteristics, it is obvious to also acknowledge the role of the user. This is important to keep in mind when thinking of the possible outcomes of decision support systems. Indeed, in our example of the job matching system, it is still up to the job seeker to engage with the proposed vacancies. Here, it has indeed been found that people's levels of extroversion and conscientiousness influence their job seeking attitudes [24].

3.5 It ain't a silver bullet

Although we argue that in the design of a DSS one should aim to incorporate these dimensions, we acknowledge that the application of this affordance reasoning is not straightforward. As with many principles, these concepts need to be operationalized and adjusted to the system's particular context. In this operationalization, one should be equally aware of the potential biases that might enter in the design. Implementing these serendipity affordances is in no way an exemption to any other biases. In this paper, we therefore want to particularly pay attention to the affordance of diversifiability. After all, the mere requirement of having a diverse (information) environment, does not sufficiently take into account existing historic biases towards certain information sources, which is due to societal power dynamics very difficult to not replicate [21]. The above-mentioned example of women getting more part-time offers is such a bias, related to the gender stereotypical role of the women as the central home and childcare provider, where paid employment is assumed of second order importance. To overcome this historical bias, we suggest that the operationalization of diversity should take into account the notion of equity as it is more adequate to deal with this type of bias.



Design for Serendipity

Diversifiability | Transferability | Sensoriability

Design for Equity

Intersectionality | Reflexivity | Power dimensions

Figure 1: Conceptual connection between design for serendipity affordances and the dimensions to enable design for equity

4. MITIGATING HISTORIC BIAS

4.1 Historic bias & equity

Any kind of decision-support system relies on one or multiple models that are trained on data. A model is an abstraction of a process which makes predictions on historic data points [40]. Consequently, the predictions and recommendations are based on data that may contain (past) societal prejudices that were (unintentionally) encoded in these historic data points. This is called historical bias, it occurs when the world as it was or is influences the model to make unwanted biased outcomes [50]. Friedman and Nissenbaum [25] coined the term "pre-existing bias" to describe the same effects. According to them, this is a bias that can be traced back to institutions, practices and attitudes. An example of an historical bias can be found in Amazon's discontinued human resources machine learning algorithm. This was used to identify possible job candidates by predicting their success based on previous employees' success. The model is no longer in use as it downgraded CVs from women due to an inferred preference for male candidates [19]. Penā Gangadharan and Niklas [41] recognise that in order to resolve discrimination it is necessary to investigate the normative practices and institutions that have contributed to this difference in treatment, as technology is created within a society's laws and practices. This means that it is necessary to examine the socio-technical dynamics of gathering data by not discounting the historical choices that created the data [41,38]. One element that is proposed to stop discrimination and to obtain equality is to design for fairness.

Historical discrimination or data provenance (the availability of data) is sometimes explicitly ignored [14] when trying to achieve fairness in models. Gilbert and Mintz [27] have made the case that in order to counter bias in data it is important to place the data within a context. This is confirmed by Binns'[8] examination of the unfairness of discrimination in relation to political philosophy's role in algormorithic bias. He concluded that fairness needs to be contextualized in order to be truly fair. He proposes to use luck egalitarianism to achieve fairness, this doctrine states that to be fairly judged a person can only bear the burden of their own choices and those burdens caused by outside influences should be taken out of the equation. He acknowledges that this contextualisation of choices is not an easy task as what is truly a choice and what is a result of the historical circumstances is in most cases hard to distinguish. Research has been done to use machine learning to discover and use causal relationships that can be found in historical data to show the bias [37]. Cardaso et al. [36] examined if it was possible to use self created biased data to validate if there was bias in the system. What these two approaches have in common is that they use data to examine or contextualize historical inequality. A shortcoming here is that this can only be established when there is data available.

Currently many of the technical solutions to bias or unfairness are focused on creating fair outcomes by algorithm [1, 2, 22]. For example, they compare if all groups are treated equally by the model. This approach, however, does not address if it is fair to the individual to be judged on the basis of a group. Moreover, the studies focused on algorithmic fairness do not address that the perception of fairness might not be the same for everyone and that the promoted fairness is not determined within a vacuum. For example, Wang et al. [52] found that fairness perception differs greatly among the 579 participants of their experiment and is not easily determined or feelings of unfairness are not easily resolved. O'Neil [40] argues that fairness is hard to grasp for models and a focus on data means that unfairness is perpetuated and often unaddressed. She [40] witnessed this in her study of police predictive modeling and these findings are echoed by the ethnographic studies of algorithms by Eubanks [23].

Data collection, or lack there off, is inherently political and is part of structural oppression [21]. One example to illustrate both the need for an historical perspective and the influence of normative practices, is in the lack of data available on the lived experience of women [42], which has impacted many fields from urban planning to health. In the context of a DSS, this kind of historical bias limits the available data and resulting interpretations and support. To remediate this, it has been suggested to design for equity [16, 17, 21].

4.2 Designing for Equity

D'ignazio and Klein [21] argue that fairness as a concept is not sufficient to address inequality. This is because fairness is judged from a current moment in time without reflection on past advantages. Therefore it would be advantageous for those that have been privileged from the start. The decisions involved in creating fairness within DSS are numerous as fairness as a concept comes in many different iterations with each determining the type of fairness that is provided to the subject of a DSS (e.g. group fairness, individual fairness, group parity and others). Determining and documenting the choice in fairness alone does not solve bias. This is echoed by Hoffmann [31] who uses anti-discrimination discourse to explain that using fairness alone does not solve bias in technological solutions as it will replicate normative structures and does not take into account the full lived experience of marginalized people. Equity on the other hand would take into account the context of a person which in turn enables an equitable outcome. The purpose of equity is to ensure that advantages and disadvantages are taken into consideration and to offer an alternative to the fairness principle. There have been initiatives to introduce what is called data justice into design practice. Data justice brings together multiple disciplines which are focused on the role of datafication within a wide variety of topics ranging from democratic procedures to the dehumanisation of decision-making [20].

When we look for inspiration on how to put this into practice, we can rely on the recent increased interest to integrate data justice in design and practice [16, 20]. The focus of these initiatives is to ensure that a plurality of voices and lived experiences are introduced within the design process, which is currently too focused on an universal experience [17]. Others have argued that integrating critical feminist theory originating from the social sciences in design of Human Computer Interaction (HCI) would ensure that there would be challenges to normative thinking [7]. According to Bardzell and Bardzell [7] critical feminist theory is uniquely qualified for this as resisting and critiquing the status quo is one of its tenets. Although the feminist theory originally set out from a gender inequality point of view, applying systematic reflexivity and a high awareness of historical power inequalities, created insight in the intersection of different inequalities coming along other social categories (e.g ethnicity, age, sexuality, ability, economical background) guiding social interaction. This lead to a body of work interested in investigating the different practices involved in creating organisations and technologies. Young [53] used their article as the basis to create a design practice for a feminist chatbot by proposing practical methods to include stakeholders within the creation process of a chatbot. Using insights from both these fields we argue to integrate the following into the diversifiablity affordance for serendipity: awareness of power dynamics, intersectionality and reflexivity.

4.2.1 Powerdynamics

Ignoring the structures of discrimination by focusing solely on single instances of blind spots disregards how these structures of power can have a wider impact [31]. D'ignazio and Klein [21] make use of the 4 domains of power (based on Patricia Hill Collins'[15] work) to explain structural oppression. First, there is the structural domain which organizes oppression via laws and policies. Second, the disciplinary domain which manages oppression by enforcing the laws and policies of the first domain. The third domain is the hegemonic domain which circulates oppression and creates acceptance via cultural activities and the media. And finally, there is the interpersonal domain which consists of the personal experience of oppression. Structural oppression can also result in a disproportionately privilege for a dominant group [21]. Robinson et al. [46] argue that disregarding structural racism while designing a system will result in it insidiously infiltrating your system on every level.

Applying these domains on the design for algorithm-decision making-processes in a labor context, there are laws and policies in place that are sensitive to equal access to jobs (structural domain of power). However, the disciplinary domain is often insufficiently equipped to enforce these laws and policies in daily practice. There are, however, cases where unequal access to jobs is challenged. It is on the third domain, the hegemonic domain, where a lot of change is still necessary. It is also in this domain a new DSS should be situated as it is a cultural expression captured in a technology: what is a good or a wrong decision. Finally, in the interpersonal domain, any oppression supported by the DSS is only shown to the person that is discriminated against. The end-users from the dominant group are not aware of their privilege, which thus remains a blind spot for them.

Designing for equity would mean to be aware of the power structures that made the decisions regarding the model and the manner of data gathering. In other words make visible who was allowed to be involved in the design process [16]. Once the power structures have been examined it becomes possible to challenge these decisions [21].

4.2.2 Complexity of intersectionality

Solely focusing on structural racism in the design of a DSS would ignore the intersectionality of most people's identities, and the related exclusion mechanisms - for example ageism, sexism, ableism, or marginalisation by wealth. The concept of intersectionality was proposed [18] to explain that race, gender and class do not operate in a silo. They intersect and on those points of intersection people's identity is constructed [16, 17]. According to Collins [15] people receive benefits and penalties based on their position on the intersections within systems of oppression. An example can be found in data sets created for training facial recognition software. Buolamwini and Gebru [12] found that the algorithm was unable to recognise black female faces because it was not only trained on a lack of black faces but it specifically lacked black female faces. There has been interest in re-examining the principles of fairness, for example, Burke [13] proposes multi-sided fairness. In this case, however, the multi-sided aspect is not considered within an individual itself as intersectionality proposes, but is related to fairness for multiple stakeholders within an algorithmic decision. Designing for equity would mean to design a system that is aware of the complex intersections of peoples' identity and to go beyond a universal design [17].

4.2.3 Reflexivity of data scientists

The third principle is reflexivity which is "the ability to reflect on and take responsibility for one's own position within the multiple, intersecting dimensions of the matrix of domination" [21, p.64]. Reflexivity is then achieved by being aware of your own position within the power dimensions and actively acknowledging the benefits and disadvantages of this position. The reflexivity achieves a transparency on the differences between data subject and data gatherer. This is important because it highlights possible gaps in the knowledge base of the data scientist.

5. IN (DESIGN) PRACTICE

On an abstract level, the trajectory of a DSS can be modelled as depicted in Figure 2. This trajectory is based on the simplified life cycle of AI as proposed by Binns and Gallo [9]. To further accommodate early consideration of ethical issues, we have elaborated the framework to reflect the necessity to consider the team who develops the AI. In addition, we have added a deeper layer to the training and test data phase to be able to differentiate between the different steps that need to be taken to procure the training and test data. A lot of the blind spots are built up in this phase, but not all as discussed earlier. Applying ethical considerations at an early stage has been found to be the most beneficial and cost effective method [25]. It is therefore important to ensure that these phases can be described as detailed as possible. We will use this framework to place the different steps for integrating equity and serendipity within all the phases of the design process. While the role of diversifiability in serendipity has been acknowledged in the design of machine learning models [35], it is rarely considered as an important aspect of the design process itself. However, what is clear from the previous discussion, is that incorporating diversifiability should not be a mere feature of the algorithm itself. Diversifiability also emerges from related design activities such as data collection or developing measures of success because they are subject of the previously described design principles for equity.

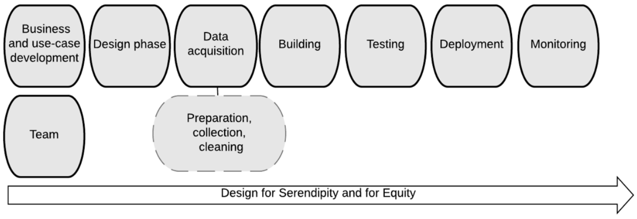

Business and use-case development
Design phase
Data acquisition
Building
Testing
Deployment
Monitoring
Team
Preparation, collection, cleaning
Design for Serendipity and for Equity

Figure 2: Based on the simplified life cycle of AI [9] an ideal typical process of the creation of a DSS is depicted, connected with the overall design for serendipity and equity as key drivers in the process

5.1 Living lab approach

We propose combining both the principles of data justice and critical feminist theory with a living lab approach. The latter is a research method that involves "multiple stakeholders, including users, in the exploration, co-creation and evaluation of (usually ICT-related) innovations within a realistic setting" [5, p. 1].

At the start of the project, in the business and use-case development phase, one assesses the stakeholders, the purpose of the DSS system and its context. Which stakeholders are under-served? A thorough stakeholder mapping should be conducted, keeping in mind the intersectionality of stakeholders' identity. In order to compensate for structural disadvantages, it will be necessary to examine where people might be penalized or privileged based on intersections of identity [15] . Moreover, it needs to be ensured that the system design is also informed by a representation (e.g. persona's) and involvement (e.g. user research, co-creation) of the most disadvantaged. In our earlier example of the job-matching system, those who have been identified as experiencing difficulties in being matched with a job should be involved in the design process. This will create additional semantic knowledge for the system to work with and consequently reduce potential biases. This could be, for example, people older than 50, educationally disadvantaged and minoritized people [51]. Involving them can be achieved by conducting user research and co-creation with representatives, not only focusing on the (historic and current) disadvantages, but also their strengths and experiences. During these sessions, system-designers should emphatically listen to the epistemic knowledge of stakeholders themselves, as they are uniquely qualified to critique the status quo [7]. Here, it is important to note that the involvement of these stakeholders needs to be a touch stone throughout the entire design process. Involving them throughout the project will ensure that their actual experiences impact the design, rather than the designers' interpretation of these experiences. This means that they would be involved in every stage of the life cycle (Figure 2). Another step where a living lab approach will benefit equity is within step 3 the data collection phase. The input is essential as interpretation of data can be tricky. Involving domain experts is essential to understand which data needs to be included to be able to make a decision. The involvement of marginalized stakeholders here would enable a new perspective on the data from a lived experience point of view. As said before the involvement of the stakeholders is throughout the lifecycle as they will be involved in the testing of the resulting model and can be involved in creating an inclusive deployment strategy.

The incorporation of equity within diversifiability can also be realized by examining the purpose of solutions together with the stakeholders identified as most impacted by the solutions. This means that the systems' beneficial impact on society is determined and possible harms for their peer group are identified. Furthermore, by interacting with these stakeholders we can learn from their experiences on how to avoid the harm and improve the solution for all. This includes an examination of the broader context by analysing the stakeholder mapping and the conducted sessions of co-creation and user research to present an overview of the possible positive and negative outcomes of the proposed solution. This can be implemented by conducting a domain analysis using classic scientific methods combined with the epistemological knowledge of the marginalized stakeholders.

5.2 Reflexivity

Next to listening and incorporating other viewpoints in the design process, there is also a need to focus on the design team. The design team should be aware of their own intersectional identities, values and position in society. Both on an individual team member level and as a group. This can be achieved by practicing reflexivity. In our example of labor mitigation, the team should first reflect on which intersection they would be placed (e.g. a team member could be a hetero-sexual 30 year old white man) and think about what this would mean for their own unconscious preferences and assumptions. Another aspect of this reflexive exercise involves studying how these intersectional identities reflect in or differ from the marginalized stakeholders. Subsequently, what kind of involvement of the marginalized stakeholders is needed to ensure that the design will have the desired impact. How are you going to realise this: an additional team member, a soundboard of experience experts, . . . Finally, if particular design choices are based on assumptions of use, these need to be disclosed in order to facilitate reflexivity on the possible impact of these assumptions later in the process. For example at the time of deployment or monitoring of the performance of the DSS. Leaflet or fact sheet approaches are commonly used as a tool to create that transparency [4, 49].

5.3 Challenges

These three principles of intersectionality, reflexivity and awareness of power dimensions help designers to ensure equity. Applying them also contributes to the diversifiability affordance that has been defined to design for serendipity. Indeed, as has been illustrated in the previous section, including diversity can only be considered as beneficial when it does not come at the cost of equality. We therefore argue that the practical methods put forward in this section should become an essential part of the design workflow of any algorithm-based decision-making system. In this way, they could be considered as complementary research approaches and practices to rather technical-oriented solutions that have been suggested before in order to deal with serendipity and diversity.

Although implementing these principles and corresponding methodologies sounds evident in theory, we acknowledge that there is still a lack of practical tools and procedural knowledge as to how to implement them in an actual design process. This is not only an open challenge related to the subjects presented in this work, but is applicable more generally to the discourse of ethical artificial intelligence. Recent works all explicitly point out that there is a need for domain appropriate ethical tools (e.g.[6, 39, 34]. From our experience as social scientists involved in the design of many of these systems, we believe that this challenge arises from (at least) two bottlenecks that need to be addressed collectively in order to be able to move forward. First of all, experience has taught us that ethical considerations do not have the attention that they should have to truly have ethics by design. They are seen as a nice to have instead of a key consideration throughout the design process. Secondly, there is a lack of a proper articulation of the actual steps that take place within the development of a DSS. This means that we do not know for sure if the trajectory in Figure 2 depicts a version of reality or is indeed solely a representation of an ideal. This hampers our ability to develop and provide more specific and concrete tools to implement the principles of serendipity and equity. We hope that the arguments in this paper demonstrate the importance of notions such as serendipity and equity in the design of algorithmic-based decision-making systems and invite scholars and practitioners from other domains to work collectively towards putting these principles into practice.

6. CONCLUSION

The availability of high volumes of data and intelligent decision-making systems present both opportunities and challenges to various actors. In the current research paradigm, the most important system objective of decision-making systems is accuracy. The relevance and applicability of the system is informally evaluated by the user and determines if they will continue to use application. While these systems allow for an enhanced informed decision-making process, the question arises to which extent the information they present is not flawed by biases and blind spots. In this work, we outlined how design principles based on serendipity and equity could help to re-mediate some of these weaknesses. The underlying rationale is that these principles will broaden the available information and semantic knowledge and in this way allow for divergent rather than convergent systems. We aimed to put forward a design rationale that would help to incorporate the principles of serendipity (diversifiability) and equity (inter- sectionality, reflexivity and power balances) in the development of DSS.

The main challenge, however, relates to the actual implementation of these methodological tools within an actual design process of a DSS. What is evident from the rationale presented in this paper, is the fact that design principles such as serendipity and equity shouldn't be limited to the activities related to the mere design or training of the system. Rather, it should be part of the entire trajectory from start to end, including testing and evaluating it continuously with several user groups. While we presented an ideal trajectory of this design and development process (cfr. Figure 2), we are aware of the fact that in reality this might not always be the case. Future research will need to be conducted to examine and elaborate on the process presented in figure 2 in order to present a realistic representation of the design of a DSS. We therefore call for an inter-disciplinary approach that considers these design principles such as serendipity and equity not merely as a nice to have, but as an essential component of the design of a DSS, and starting from this, discusses how these practices can be met in the actual design of these systems.

APPENDIX

A. ADDITIONAL AUTHORS

Additional authors: Pieter Ballon(imec-SMIT-VUB, email: Pieter.Ballon@vub.be)

B. REFERENCES

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.

[2] Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN.

[3] Al-Otaibi, S. T., & Ykhlef, M. (2012). A survey of job recommender systems. International Journal of the Physical Sciences, 7(29), 5127-5142.

[4] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Reimer, D. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

[5] Ballon, P., Van Hoed, M., & Schuurman, D. (2018). The effectiveness of involving users in digital innovation: Measuring the impact of living labs. Telematics and Informatics, 35(5), 1201-1214.

[6] Ballon, P., Duysburgh, P., Fanni, R., Franck, G., Heyman, R., & Laenens, W. (2019). D2.2. Raamwerk voor ethische validatie van AI. Kenniscentrum Data & Maatschappij, Brussel, België (Authors alphabetically placed)

[7] Bardzell, S., & Bardzell, J. (2011, May). Towards a feminist HCI methodology: social science, feminism, and HCI. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 675-684).

[8] Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. arXiv preprint arXiv:1712.03586.

[9] Binns, R., & Gallo, V. (2019). An overview of the Auditing Framework for Artificial Intelligence and its core components. Retrieved February 17, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

[10] Björneborn, L. (2017). Three key affordances for serendipity: Toward a framework connecting environmental and personal factors in serendipitous encounters. Journal of Documentation, 73(5), 1053-1081. 10.1108/JD-07-2016-0097

[11] Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and information technology, 15(3), 209-227.

[12] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91).

[13] Burke, R. (2017). Multisided fairness for recommendation. arXiv preprint arXiv:1707.00093.

[14] Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. In Advances in Neural Information Processing Systems (pp. 3539-3550)

[15] Collins, P. H. (2002). Black feminist thought: Knowledge, consciousness, and the politics of empowerment. Routledge.

[16] Costanza-Chock, S. (2018). Design Justice: towards an intersectional feminist framework for design theory and practice. Proceedings of the Design Research Society.

[17] Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. Journal of Design and Science. 10.21428/96c8d426

[18] Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stan. L. Rev., 43, 1241.

[19] Dastin, J. (2018) "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters

[20] Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring data justice: Conceptions, applications and directions.

[21] D'Ignazio, Catherine, and Lauren F. Klein. Data feminism. MIT Press, 2020.

[22] Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2017). Decoupled classifiers for fair and efficient machine learning. arXiv preprint arXiv:1707.06613.

[23] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[24] Fort, I., Pacaud, C., & Gilles, P. Y. (2015). Job search intention, theory of planned behavior, personality and job search experience. International Journal for Educational and Vocational Guidance, 15(1), 57-74.

[25] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.

[26] Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10, 257. 10.1145/1864708.1864761

[27] Gilbert, T. K., & Mintz, Y. (2019, January). Epistemic Therapy for Bias in Automated Decision-Making. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 61-67).

[28] Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. https://www. technologyreview. com/s/612876/this-is-how-ai-bias-really-happensand-whyits-so-hard-to-fix.

[29] Ho, J., & Tang, R. (2001, September). Towards an optimal resolution to information overload: an infomediary approach. In Proceedings of the 2001 international ACM SIGGROUP conference on supporting group work (pp. 91-96).

[30] Holone, H. (2016). The filter bubble and its effect on online personal health information. Croatian medical journal, 57(3), 298

[31] Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900-915.

[32] Johnson, S., & From, W. G. I. C. (2010). The Natural History of Innovation.

[33] Kaminskas, M., & Bridge, D. (2016). Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. ACM Transactions on Interactive Intelligent Systems, 7(1), 1–42. 10.1145/2926720

[34] Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. Big Data & Society. 10.1177/2053951720915939

[35] Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. Knowledge-Based Systems, 111, 180-192.

[36] L. Cardoso, R., Meira Jr, W., Almeida, V., & J. Zaki, M. (2019, January). A framework for benchmarking discrimination-aware models in machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 437-444).

[37] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019, January). Fairness through causal awareness: Learning causal latent-variable models for biased data. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 349-358).

[38] Milan, Stefania, and Emiliano Treré. 2019. “Big Data from the South(s): Beyond Data Universalism.” Television & New Media 20 (4): 319–35.

[39] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics, 1-28.

[40] O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

[41] Peña Gangadharan, S., & Niklas, J. (2019). Decentering technology in discourse on discrimination. Information, Communication & Society, 22(7), 882-899.

[42] Perez, C. C. (2019). Invisible Women: Exposing data bias in a world designed for men. Random House.

[43] Race, T. M., & Makri, S. (Eds.). (2016). Accidental information discovery: cultivating serendipity in the digital age. Elsevier

[44] Reviglio, U. (2019a). Serendipity as an emerging design principle of the infosphere: challenges and opportunities. Ethics and Information Technology, 21(2), 151-166.

[45] Reviglio, U. (2019b). Towards a Taxonomy for Designing Serendipity in Personalized News Feeds. http://informationr.net/ir/24-4/colis/colis1943.html

[46] Robinson, W R, A Renson, and A I Naimi. 2020. “Teaching Yourself about Structural Racism Will Improve Your Machine Learning.” https://academic.oup.com/biostatistics/article-abstract/21/2/339/5631851.

[47] Sadler, E. B., & Given, L. M. (2007). Affordance theory: a framework for graduate students' information behavior. Journal of documentation, 63(1), 115-141.

[48] Smets, A., Walravens, N. & Ballon, P. (2020). Designing Recommender Systems for the Common Good. In Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy. ACM, New York, NY, USA, 3 pages. 10.1145/3386392.3399570

[49] Sokol, K., & Flach, P. (2020, January). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 56-67).

[50] Suresh, H., & Guttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. ArXiv, abs/1901.10002.

[51] Van Humbeeck, G. (2020, April). AI VDAB. Presentation presented during Data Date 2, Kenniscentrum Data & Maatschappij. https://data-en-maatschappij.ai/nieuws/data-date-2-ai-en-rekrutering

[52] Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

[53] Young, J. 2017. “Designing Feminist Chatbots” 2017. https://www.ellpha.com/list/2017/9/23/designing-feminist-chatbots.