

Human-Centred Artificial Intelligence

A Contextual Morality Perspective

van Berkel, Niels; Tag, Benjamin; Goncalves, Jorge; Hosio, Simo

Published in:
Behaviour and Information Technology

DOI (link to publication from Publisher):
[10.1080/0144929X.2020.1818828](https://doi.org/10.1080/0144929X.2020.1818828)

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

[Link to publication from Aalborg University](#)

Citation for published version (APA):
van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2020). Human-Centred Artificial Intelligence: A Contextual Morality Perspective. *Behaviour and Information Technology*. <https://doi.org/10.1080/0144929X.2020.1818828>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Human-Centred Artificial Intelligence: A Contextual Morality Perspective

Niels van Berkel^a and Benjamin Tag^b and Jorge Goncalves^b and Simo Hosio^c

^aAalborg University, Aalborg, Denmark; ^bThe University of Melbourne, Melbourne, Australia;

^cUniversity of Oulu, Oulu, Finland

ABSTRACT

The emergence of big data combined with the technical developments in Artificial Intelligence has enabled novel opportunities for autonomous and continuous decision-support. While initial work has begun to explore how human morality can inform the decision-making of future Artificial Intelligence applications, these approaches typically consider human morals as static and immutable. In this work, we present an initial exploration of the effect of context on human morality from a Utilitarian perspective. Through an online narrative transportation study, in which participants are primed with either a positive story, a negative story, or a control condition ($N = 82$), we collect participants' perceptions on technology that has to deal with moral judgment in changing contexts. Based on an in-depth qualitative analysis of participant responses, we contrast participant perceptions to related work on Fairness, Accountability, and Transparency. Our work highlights the importance of contextual morality for Artificial Intelligence and identifies opportunities for future work through a FACT-based (Fairness, Accountability, Context, and Transparency) perspective.

KEYWORDS

Contextual morality; Algorithmic bias; Artificial intelligence; Ethics; Decision-support

1. Introduction

Artificial Intelligence (AI) algorithms are increasingly deployed in the real world, covering both one-off, high-stakes decision support as well as mass scale daily routine assistance. Examples include the use of AI in domains such as justice [25], transportation [11], and health [22, 15] – as well as day-to-day interactions such as location-based activity recommendation systems [89] and the selection of personalized news items [80]. Although AI-technology has the potential to support and augment human-decision making, the negative consequences of such automated technologies have recently covered the front pages of news outlets, reporting on racism and sexism [90, 18, 60], as well as other biases in recent AI deployments [49]. These concerns have recently led to a political interest in this area. At the recent G20 summit, an economic council of the world's wealthiest nations, ministers agreed on the need for human-centric AI principles [30]. Given the potentially far-reaching consequences and the (typically) opaque nature of 'AI in the wild' systems, recent work has begun to explore critical aspects such as the Fairness, Accountability, and Transparency (FAT*) of AI [13, 15, 25, 84, 86] – resulting

in the creation of guidelines for the design and development of AI applications. In a recent review of AI these emerging guidelines, Jobin *et al.* compare nationwide and international (*e.g.*, European Union, IEEE) AI principles [52], highlighting a number of major differences in the interpretation of FAT and other critical aspects, as well as the issues, domains, and actors targeted by the guidelines.

Although work on FAT* is critical, the review by Jobin *et al.* highlights that these factors cannot be seen in separation from the moral values and beliefs which drive human decision making [19, 52]. As human-labelled data is routinely used to inform a system’s reasoning, any AI-system will, by definition, be biased to the (current) moral standards of those responsible for contributing (‘ground truth’) data. Previous work indicates that moral standards differ between people, context, and culture [45, 82, 7]. Hofstede suggests that national cultures can be described along six dimensions (Power Distance, Uncertainty Avoidance, Individualism/Collectivism, Masculinity/Femininity, Long/Short Term Orientation, and Indulgence/Restraint), which explain differences in cultural norms [45]. For example, Hofstede found that the United States scores highest in Individualism, followed by other Western countries, whereas the majority of Eastern countries score higher on Collectivism. Similarly, Awad *et al.* find significant differences between participants originating from Western, Eastern, and Southern clusters [7]. However, AI systems typically contain human-labeled data originating from unrepresentative and uniform (*e.g.*, cultural background, age, location) populations (*e.g.*, US crowdsourcers working at Amazon Mechanical Turk) – consequently biasing the behaviour of AI systems [25, 43]. Furthermore, while current AI-powered systems are primarily based on a combination of historical and human-labelled data, future AI systems will inevitably have to function in previously unencountered scenarios. As such, their reasoning should not only consider the ‘optimal’ end-result but also acceptable ways to achieve this goal (or decide that the goal should be abandoned altogether). Therefore, future AIs will have to autonomously reason about ‘right’ and ‘wrong’ to inform their decision making and ensure transparency towards end-users and regulatory oversight.

Initial work has begun to explore how human morality, the codes of conduct that distinguish right and wrong behaviour [33], can inform the decision-making of future AIs [7, 11]. However, these approaches typically consider human morality as a static concept [73]. In contrast to this work, the Philosophy and Psychology literature highlights that our morality is anything but constant and that our moral compass is profoundly affected by our context [59, 83]. Indeed, identifying what is considered as ‘right’ and ‘wrong’ is heavily dependent on the situation in Utilitarianism [65], Deontologism [12], and Virtue ethics [36]. To align with the ethical perspective of end-users, a distinct subset of AI applications should, therefore, be capable of considering the user’s context in their decision-making process. This is in contrast to relying on a constant set of guidelines, typically following a Silicon Valley-imposed mindset, which are not necessarily aligned with the user’s viewpoint. We, therefore, argue that an essential element for future AIs is to exploit contextual information in order to determine the expected moral behaviour and adjust their actions accordingly.

In this work, we explore the role of morality in AI-based decision-making, pointing our attention to the under-explored concept in Computer Science of *contextual morality*. Following an extensive review of the FAT* and ethical AI literature, we introduce the Contextual Morality Framework. The Contextual Morality Framework suggests that in order to increase the fairness, accountability, and transparency of future AI-powered systems, we must carefully consider the context within which the system is assessed. Therefore, a true integration of context into FAT research, *i.e.* FACT, is required.

Through an elicitation study, in which we apply the ‘narrative transportation’ method, we collected rich, open-ended feedback on the participants’ perceptions of the issue of developing technology that has to deal with moral judgment. Through a between-subject design, we engage participants with a text-based stimulus and employ validated questionnaires to gauge their immersion to the presented stimulus and moral values. Our results shed light on the topic of morality in the development of AI-based agents. Building an understanding of the (end-)users of technology and their expectations plays a pivotal role in supporting researchers and developers. The findings in this paper contribute to the intellectual debate on ethical AI and initiate a research perspective for context-sensitive understanding in future AI applications.

2. Related Work

The study of morality has a rich history, with an extensive array of viewpoints developed since the ancient philosophers. Works by Aristotle and others focused strongly on ‘virtue ethics’, *i.e.*, characteristics which make for a ‘good’ person [4]. Some of the virtues advocated by Aristotle include ‘courage’, ‘truthfulness’, and ‘modesty’. However, the implementation of virtues in AI applications is inherently challenging; *e.g.*, how would one go about quantifying a courageous algorithm through a rule-based approach? Howard & Dungan argue that, through the use of soft-computing methods such as neural networks and evolutionary computation, moral agents can obtain a certain level of autonomy which allows them to make decisions outside of a set of predefined rules [48].

A contrasting perspective that has seen initial uptake is the use of a Utilitarian perspective. Utilitarianism states that ethical decision making maximizes value (*i.e.*, utility) [65]. As such, an algorithm can be constructed that - as long as the utility of potential outcomes is defined - can calculate the morally most optimal decision. Bonnefon *et al.* and Awad *et al.* utilize this perspective in studying the perception of the moral implications of autonomous vehicles (AVs) [7, 11]. Their work builds on the premise that AV algorithms will eventually have to make decisions which negatively affect either the passenger(s) or other road users. Participants were asked to make moral judgments on a range of decision examples, with the results of the study offering implications for the implementation of both AV algorithms and policy.

However, despite offering an excellent example of quantifying morals on a large scale, this work fails to take into account recent advances in Philosophy which reveal that a person’s moral judgment is variable. For example, Leavitt *et al.* showed that one’s assumed professional identity (*e.g.*, manager vs engineer, soldier vs medic) substantially influences moral judgment [59]. Similarly, Reynolds *et al.* show how contextual cues can shape ethical behaviour in a business setting [73]. These examples demonstrate that both the long-term and short-term context affect the moral judgment of individuals. As such, we argue that AI applications should consider the context in which they operate in order to function successfully. A relevant theory from social psychology is that of situationism, which describes that it is not a person’s character traits, moral beliefs, or past behaviour which can be used to forecast their behaviour but rather the characteristics of the situation they are in [55]. Harman states that “*empirical studies designed to test whether people behave differently in ways that might reflect their having different character traits have failed to find relevant differences*” [42], arguing that character-based virtue ethics should be abandoned. The notion that researchers attribute behaviour to an individual’s character rather than the situation (*i.e.*, context) which they are in, has been called

the ‘attribution error’. In response to this critique, Kamtekar argues that the character traits considered in the situationist perspective differ from the way in which character is interpreted in virtue ethics. Whereas the empirical studies in situationism consider character traits independently, Kamtekar states that “*the conception of character in virtue ethics is holistic and inclusive of how we reason*”. While the debate between situationism and virtue ethics is not the main focus of this paper, we acknowledge that our argument follows a situationist perspective and identify this as an important underpinning of the remainder of the paper.

2.1. *Technology and Morality: Fairness, Accountability, Transparency, and Ethics*

Automated algorithms have long been understood as being part of the realm of *means*, whereas morals are said to be located in the realm of *ends* [58]. With the evolution of automated systems to autonomous systems, however, this clear distinction has started to blur. Autonomous systems are, by definition, self-serving and, therefore, occupy the ends realm as well as the means realm [27]. The ever-increasing impact of AI-powered applications on daily human life has led to an intensified investigation of potential applications of these new algorithms. This work has solidified itself under the umbrella of the following terms; Fairness, Accountability, and Transparency.

2.1.1. *Fairness*

Fair AI aims for decision making which is equitable between different classes (*e.g.*, demographic differences such as age). Reasons for unfair decision making primarily stem from biases in the dataset used to train a classifier. Barocas and Selbst distinguish two ways in which biased training data can result in discriminatory models [8]. First, the training data can consist of cases in which one group was favoured or disfavoured over other groups. By training a decision-making model on this data, the final model may repeat biases from the past as training data is considered to consist of valid cases. Second, training data can be biased towards a specific sample of the population. Even if the dataset is entirely accurate, by under- or over-representing certain samples of the population, biases can emerge. Liu *et al.* show that common criteria for fairness may not be sufficient to give equal opportunity across different demographic groups when it comes to selecting candidates for certain opportunities (*e.g.*, job selection) in score-based ML algorithms [63]. Common fairness criteria actually tend to cause delayed disadvantages for underrepresented groups, thus, Liu *et al.* recommend to consider the long-term outcomes when building fair AI systems. An approach to tackle this issue at the root has been taken by Albarghouthi and Vinitzky, who propose Fairness-Aware Programming as a solution to include rules for fairness in the code – automatically monitored by a runtime system which checks for, and subsequently reports, rule violations [2].

2.1.2. *Accountability*

The automated, and often obscure, decision-making processes employed by AI applications raise questions regarding their **accountability**. Diakopoulos discusses necessary measures to guarantee clearer levels of accountability and distinguishes between requirements for the private and the public sector [24]. Diakopoulos argues that appropriate mechanisms have to be put in place in all domains where autonomous systems create

output. An example is the use of accountability reports in journalism, which enable sample testing for biases and hoaxes in automatically generated texts. These inherent biases can be results of the human engineers’ own biases or because they are hidden in the training data used to train ML algorithms. ML algorithms that rely on big training data sets used to predict human behaviour, risk propensity, or health status are often called ‘Black Boxes’ [70], as they raise questions not only with regards to accountability but also in respect to the system’s transparency. Research has shown that uncovering black boxes is effective in making ML more understandable, *e.g.*, enabling children to actively engage with simple ML mechanisms in their everyday life [44].

2.1.3. Transparency

Work on the **transparency of AI** that aims to increase the ability for people to understand the reasoning behind AI-based decision-making processes. This includes, *inter alia*, work on interpretable models such as General Additive Modelling [15]. In essence, these models support high intelligibility through reduced complexity – often only at a small cost of reduced accuracy. A critical aspect of a transparent system is the ability to explain how the system arrived at a given decision. Previous works highlight the field of Human-Computer Interaction (HCI) as a potential key contributor to explainable AI (XAI) research: “*Given HCI’s focus on technology that benefits people, we, as a community, should take the lead to ensure that new intelligent systems are transparent from the ground up*” [1].

Veale *et al.* have discovered deep discrepancies between the awareness of a need for more transparency, fairness, and accountability and the actual algorithms developed and put to work in the public sector [84]. They argue that the HCI community has a potential stake in supporting fair and transparent execution of AI systems by governmental power. Based on these findings, Holstein *et al.* indicate future directions for HCI research to support ML engineers [46]. A major issue identified in relation to increased transparency is a change in user behaviour as the result of their understanding of the algorithm and its functionality. Research has shown that users of the Yelp recommendation platform¹, after learning about the functionality of the algorithm, will either abandon the platform or start writing specifically for the algorithm [29]. Another study on the lack of transparency of the Airbnb² evaluation algorithm has identified anxiety amongst accommodation hosts [51]. A solution to increase understanding of autonomous AI systems has been offered by Iyer *et al.*, who demonstrate that object saliency map visualizations, *i.e.*, visual explanations of why an action was taken by the system, support intelligibility of the decisions of the system by the user [50]. A different approach can be found in the uptake of accompanying datasets with ‘datasheets’ to improve transparency and accountability, as is already a standard practice in the electronics industry [32]. IBM researchers proposed a solution called ‘Supplier’s Declaration of Conformity’ (SDoC), which describes a fact-sheet providing information about four key aspects identified by IBM as necessary for developing trusted AI, namely Fairness, Explainability, Robustness, and Lineage. The SDoC is provided by the AI service provider, and contains information and explanations on the purpose, performance, safety, security, and provenance of AI systems, enabling users to understand and examine these AI systems both prior to and during their deployment [6].

¹Yelp website at <https://www.yelp.com/>

²Airbnb website at <https://www.airbnb.com>

2.1.4. Ethics

The in-the-wild deployment of AI systems will inevitably result in scenarios in which systems have to choose between two potentially adverse outcomes. A practical and well-known example is that of an autonomous vehicle (AV) which has to choose between harming the well-being of passengers in the vehicle or that of people outside of the car. Bonnefon *et al.* investigated the responses of people to these social dilemmas and identified the complexity of the problem: participants preferred to be transported by AVs which protect the passengers, but simultaneously want other AVs to react in a utilitarian way, *i.e.*, potentially sacrificing the well-being of the passengers [11]. This complicated and paradoxical situation has led to the crucial discussion on how to make autonomous system decision-making (more) **ethical**. Savulescu and Malsen state that “*an AI that could monitor, prompt and advise on moral behaviour could help human agents overcome some of their inherent limitations*” [76]. The authors argue that by monitoring the factors which affect a person’s moral decision making, the AI can make individuals aware of their (contextual) biases and recommend the ‘right’ decision path. Chopra *et al.* argue that moral decision-making cannot merely be explained through ethical dilemmas, but is instead a context-dependent process within which the autonomous agent acts [17]. Expanding the analytic focus and including contextual factors in the analysis of ethical decision-making processes would thus enable researchers to develop more dynamic models of moral decision making [87, 31]. As such, we argue not for solely considering what is ethical or unethical, but rather for assessing how these judgements may differ between different contexts (*e.g.*, cultures, activities, moods).

2.2. Narrative transportation

The narrative transportation theory states that due to the immersion in a story (whether presented through text, video, or other stimuli), participants change their attitudes and intentions to reflect the story’s circumstances [37, 83]. Narrative transportation has been employed in a number of different domains, primarily in the area of Communication and Psychology. For example, Morgan *et al.* investigate the effect of popular television dramas on the attitude, behaviour, and knowledge of viewers in relation to organ donation [66]. Their research reveals that viewers are more likely to become an organ donor if the TV show explicitly encouraged donation and discusses the potential benefits of donation. Similarly, an increasing number of research papers suggest that smoking displayed in movies increases smoking behaviour among adolescents [67].

In contrast with the aforementioned studies which investigate existing narratives (*e.g.*, TV shows), researchers have also applied narrative transportation as a method of scientific inquiry. In a study investigating prejudices, Johnson *et al.* evaluated the effect of a narrative text describing Arab-Muslim culture containing a number of counter-stereotypical exemplars [54]. Following this, participants’ prejudice against Arab-Muslims was measured using an Implicit Association Test (IAT) [39]. The results show that both implicit and explicit prejudice decreased among participants which read the full narrative as compared to those presented with a condensed narrative or no narrative at all. Furthermore, participants report increased empathy for Arab-Muslims. Similar results were reported by Johnson *et al.* on the perception of different races [53].

A study by Grizzard *et al.* revealed that playing a game in a guilt-inducing condition results in increased moral sensitivity [40]. Participants played the game either in the role of terrorist (guilt inducing) or the role of a UN soldier (control condition). Participants’ level of guilt increased significantly when engaging in unjustified violence in the video

game. The aforementioned examples indicate that it is possible for people’s moral standards and convictions to change in a relatively short period of time induced by a transporative experience.

3. The Contextual Morality Framework

Although the interplay between technological developments and human values has been widely established in the literature (see *e.g.* [57]), this notion falls short in recognising that values differ from person to person and are far from constant. Context has a large effect on how we judge events or actions, as for example studied in narrative transportation studies [54, 53, 66]. In this paper, we propose a framework describing contextual morality and its relationship to Artificial Intelligence applications. Despite the fact that AI applications face increased scrutiny with regards to their (perceived) fairness and accountability, current work on AI has not significantly considered the effect of context on user behaviour and expectations. Here, we consider morality from a descriptive perspective – *i.e.*, “*certain codes of conduct put forward by a society or a group (such as a religion), or accepted by an individual for her own behavior*” [33]. We subsequently define contextual morality as the moral convictions of an individual or group under a specific circumstance, emphasising that these convictions may change following an adjustment of the circumstances. As such, we argue that a society’s ‘code of conduct’ typically reflects a certain level of plasticity which needs to be reflected in the technological artefacts we design for society.

3.1. Relation to Fairness, Accountability, and Transparency

Fairness, Accountability, and Transparency (frequently abbreviated as FAT*) have emerged as important considerations in AI. Lepri *et al.* [61] define fairness as “*the lack of discrimination or bias in decision making*”. The authors identify a diverse range of perspectives on how fairness can be achieved, which include among others ‘group fairness’ (each group in the dataset should receive an equal fraction of each possible outcome) [14], ‘individual fairness’ (similar people should be treated similarly) [26], and ‘equality of opportunity’ (people of equal talent and motivation should be offered the same perspectives regardless of their place in the current social system) [41]. Accountability is defined as the “*assumption of accepting the responsibility for actions and decisions*” [61]. Although the literature describes a number of ways to achieve accountability (see *e.g.* [56, 84]), transparency is often seen as a key enabler of accountability in AI. Lepri *et al.*’s definition of transparency describes it as the “*openness and communication of both the data being analyzed and the mechanisms underlying the models*” [61]. Transparency of a model can be achieved at a number of levels, most notably at the level of the model (*i.e.*, grasping the model’s workings), at a component level (*i.e.*, parameters and computation can be intuitively explained), or at the algorithmic level [61]. Lack of either one of the aforementioned factors can severely hamper a user’s understanding of the inner workings of an application [28].

Although Fairness, Accountability, and Transparency are all critical to the adaptation of trustworthy AI algorithms, we note that none of these concepts consider the variability with which individual’s perceive and evaluate the world around them. As such, a system which is considered fair in one context may be evaluated considerably different in a different context. In order to evaluate an AI application, it is therefore critical to consider the contextual factors under which the algorithm is assessed and deployed.

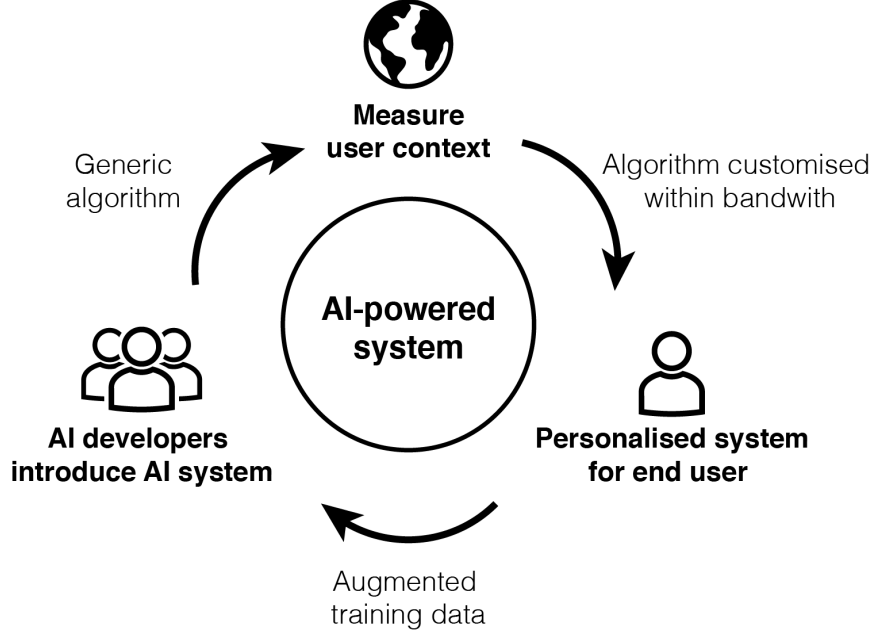


Figure 1. Conceptual diagram of the continuously evolving interaction between AI development and user context - as captured in the Contextual Morality Framework.

Given the virtually endless possibilities in which a person’s context can affect their morality, we propose a continuously evolving system describing the interplay between user context and AI behaviour (see Figure 1). As the user’s context shifts, AI-powered systems can – within the limits of a predefined bandwidth of solutions – identify a solution most optimal to the user’s current contextual morality.

3.2. Studying Contextual Morality

Studying contextual morality requires us to clearly identify which aspects of context we are capturing. Dey defines context as “*any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves*” [23]. Following this widely used definition, we stress that context is an immensely complex set of configurations and scenarios. Despite these complexities, research in the field of HCI and UbiComp has highlighted that it is possible to study and evaluate the effect of a subset of contextual variables.

Given the initial stage of research into human-AI interaction, we therefore propose that the initial focus of work in this area focuses on contextual variables previously identified in the morality literature. Although the effect of these variables, such as a person’s cultural background [45, 85] or their current company (*i.e.*, alone versus together) [88], on a person’s current moral stance have been identified – how to capture and model these variables in AI systems is currently unexplored. By following this pragmatic approach, researchers are able to isolate individual contextual variables for systematic analysis.

4. Elicitation study

We employ an online narrative transportation study, allowing for a scalable and structured data collection approach. According to the narrative transportation theory, participants engaged in a story change their attitudes and intentions in order to reflect the circumstances of a given story [83]. In order to verify whether this theory can be applied to the investigation of contextual morality, we employ a total of three conditions across a between-subjects design. In two of the conditions, participants will be presented with a fable. Fables are short stories, typically containing a clearly identifiable moral or ‘life lesson’. In this study we follow Samuel Croxall’s version [20] of two Aesop fables (‘The Lion and the Mouse’ and ‘The Wolf and The Lamb’), with an illustration of these stories by Milo Winter³. Finally, we include a control condition in which participants are not shown any primer.

Condition	Title	Perry Index*	Wordcount
Fable - Positive	The Lion and the Mouse [20]	150	207
Fable - Negative	The Wolf and The Lamb [20]	155	266
Control	-	-	-

Table 1. Overview of study conditions. *The Perry Index is a commonly used indexing system to refer to Aesop fables.

Following presentation of the stimuli, participants answer a set of questionnaires;

- First, we collect basic demographic data of the participant, including their age and gender (allowing for open text input).
- Second, we measure the participants’ level of immersion using the Transportation Scale - Short Form (TS-SF) [3] (not included in the control condition). TS-SF is a validated questionnaire which consists of six question items. Each item is rated on a 7-point Likert scale, ranging from ‘1: Not at all’ to ‘7: Very much’. Two of the six items assess imagery for two main characters. We ensured that each of the selected fables contains two main characters. The purpose of including the TS-SF is to verify whether the chosen stimuli are a) able to elicit narrative transportation, and b) comparable in terms of their narrative transportation qualities. Furthermore, we included a verification question (also known as ‘golden question’) to test whether the participant paid sufficient attention while shown the stimuli. For the Fable conditions, this question asked where the two main characters of the story had met, offering participants the choice between four options.
- Third, we obtain a measurement of the participant’s current moral standards using the Moral Identity Questionnaire (MIQ) [10]. The MIQ is an empirically validated, 20-item questionnaire, assessing both the salience of moral integrity and moral self [10]. Questions were presented in randomized order.
- Finally, and most crucial in our data collection, we ask participants to reflect on the effect of context on their values and behaviour. We ask participants to state whether they believe that moral values change depending on their context (‘yes’, ‘no’, ‘not sure’), and ask participants to elaborate on their answer in a

³Text and images are in the public domain and can be found on the Project Gutenberg website at <https://www.gutenberg.org>

free-text input field. Additionally, participants are asked to reflect on the topic of automated decision making in intelligent (mobile) systems through free-text input – and are given a few examples to ignite their thought process. In order to obtain rich insights, we offered a small monetary incentive to those participants who provided particularly insightful responses.

We deploy both the stories and questionnaires online and recruit participants using Prolific⁴ allowing us to collect responses from verified participants and deploy our survey at scale. Participants were compensated at an hourly rate of 10 GBP, with bonuses provided manually to 20% of participants. Our compensation structure exceeds minimum wage in both the UK and the US (respectively 8.21 GBP and 5.70 GBP at the time of writing). Participants could participate only once in this study.

5. Results

A total of 62 participants took part in our study, with a total of 21 participants for the positive fable condition, 20 for the negative fable condition, and 21 for the baseline condition. Participants had an average age of 32.3 years old ($SD = 11.1$). 38 of our participants identified as female, 24 participants identified as male. The first language of all of our participants is English. A total of three participants failed to answer the verification question correctly and are therefore excluded from our data analysis. As a result, we end up with 20, 18, and 21 participants for the positive fable, negative fable, and baseline condition respectively. We first present an in-depth assessment of our qualitative results, which informs the subsequent quantitative analysis.

5.1. Thematic Coding

We performed thematic coding on the participant’s responses. In doing so, we followed a three step process. First, we read all of the participants’ responses to obtain a global overview of the differences and similarities found within the dataset. Second, each of the three coders annotated each individual participant response following an inductive coding approach. Subsequent discussion between the coders resulted in three overarching themes. The themes, ‘moral influencers’, ‘moral-infused technology’, and ‘setting moral standards’, highlight the concerns, suggestions, and mindsets brought forward by our participants.

5.1.1. Moral influencers

Although only half of the participants expressed the belief that their context affected their moral values, those who did were often able to provide clear examples of the factors affecting them in their daily life. The most commonly used explanation for changing moral values was the effect of other people. “*I believe that when I go to my classes, my moral values are more secure and refined but when I am hanging out with my friends, I throw moral values out the window.*” (P26). Participants also revealed how the company of others may affect their intentions, even if it is simply to portray a ‘better’ version of themselves; “*Being around certain groups and individuals is more likely to make myself check my wallet for change (even if I know I don’t have any on me) when around rough sleepers. If I am on my own and know I do not have change, I*

⁴Prolific website at <https://prolific.ac/>

am even less likely to check. This leads me to believe my values (at least in this area) are sometimes more about performance than actual altruism.” (P02).

Several participants note that it is not solely the effect of individuals or small groups (e.g. family, friends) which affect their moral values, but that the culture by which they are surrounded has a profound impact as well. *“I believe that moral values can vary depending on the culture. A certain action or belief may be considered moral in one place, but not in another. Depending on the culture that I am surrounded by, my actions may take a more conservative (safe) turn than they otherwise would if I was by myself or around similar minded people.”* (P60). Given the increased multicultural interactions people have in their lives, this notion is increasingly relevant and currently under-explored. Our participants note that even if their personal values may not necessarily align with that of their current peers, there is a strong pressure to go along with the moral stance of the group; *“I believe group mentality plays a big part in ‘moral values’ and that while you may not act on these new values, you can often just go with the group.”* (P54).

Finally, participants seem to have a clear understanding of their actions on their moral decision making. *“After a glass or two of wine, it’s easier to make a decision which would feel wrong otherwise.”* (P63). A similar sentiment is found in the reflection on the effect of context on participant’s morality; *“At work, I need to be in a supervisor role so would do everything by the book. At home, I can be more relaxed as my family knows me.”* (P57).

As aforementioned, about half of our participants do not believe that context affects their moral values. These participants express the belief that their morality is either immutable, or not affected by any short-term changes in their environment. This stand-fast position is expressed by participants along the following lines; *“My morals were decided by myself a long time ago and there’s very little in life that could happen that would make me change.”* (P14). Several participants report a middle ground, where they acknowledge changes in their behaviour as the result of a changing context, but do not necessarily believe that their values adjust accordingly; *“Your values should be set things (potentially changing over a long period of time, or after learning new information). How you express them may vary depending on context etc., but the values themselves shouldn’t change.”* (P06).

5.1.2. Moral-infused technology

When confronted with the moral aspects of intelligent (automated) support-systems, a large number of participants raised concerns. Although some participants expressed a more general resentment; *“Allowing a computer to think for us is encouragement to not think for ourselves. I think it’s wrong!”* (P14), most participants were able to articulate specific concerns.

Part of these concerns are based on the participant’s understanding of how such AI-powered technology may work; *“Context plays a part in morality, and it would be difficult to program this understanding within a computer system.”* (P07). A number of participants further extend this argument by proposing that morality is something that distinguishes humans from machines; *“I think this is a difficult question to answer because I am not sure it is possible or if I am even comfortable with a computer (in effect) taken on traits of my morality. As a person, it actually makes me cringe to think that a computer feels it could understand me but maybe that is because I am not very comfortable with the depths AI can go to beyond say Spotify suggestions.”* (P16). In addition, participants suggest that concerns they may have in one application area

of AI may not necessary be relevant to other applications or services; *“In something as seemingly trivial as text prediction then I think the algorithm should try do that job as well as it can. [...] It gets much tougher when other decisions are being made - for instance a dating app.”* (P46).

Other participants expressed the fact that morality differs both between and within persons, impeding moral factors in technology. For example, P07 refers to cultural and family differences that shape our moral viewpoints; *“Different cultures teach different moralities, and even different families teach different ideas. This would make it near impossible for a computer system to accurately suggest things that would be considered morally just to everyone who utilizes the system”*. Several participants note that they would be more like to trust technology if it was in line with their personal contextual morality; *“I would be more likely to use such technology and rely on it more heavily if it was closely in tune with my personal, contextual morality.”* (P60). One of the aspects repeated by participants is the potential effect of their mood; *“Yes contextual morality is important, as our decisions sometimes alter depending on different circumstances, such as moods and how we are treated. Different types of people enhance or downplay some aspects of our personalities and this would mean that a mobile system would need to have flexibility in their thought and decisions to reflect this varying personality.”* (P53).

5.1.3. Setting moral standards

Finally, when it comes to deciding on the moral standards embedded in everyday technological systems, participants provided a number of perspectives. As expected, several participants note that moral standards and behaviour should be based on past user behaviour. These participants note that moral values are highly personal, and as such do not consider the use of identical moral rulesets across all users as acceptable; *“I would expect any such system to be trained by my own usage, as such it would match my moral values. [...] Morals are highly personal, and it would feel improper to force my morals on another, or to have theirs forced upon me.”* (P06).

An alternative approach proposed by participants is to consider a group of trustworthy people to devise a set of moral standards for AI-based technology. Participants in favour of this approach stressed that this group should consist of people outside of the commercial companies or institutions responsible for creating these algorithms; *“Moral people who have proved them self in society should be involved with this. Such as volunteers, teachers, police, nurses. It should not be left to the owners of these companies who often have to break their morals to get where they are in life”* (P15). In addition, participants note that such a group should be sufficiently diverse in order to obtain a wide range of viewpoints; *“A panel of well informed, educated range of scientists. I think a range of experts from a variety of backgrounds would reach a good consensus on what is morally right.”* (P17). *“Context should always be included when using intelligent systems. People who choose the moral values of the system should come from a varied group of people; different people have different morals and therefore what one person may think as immoral, one may not think that.”* (P25). Finally, one participant highlights a grassroot approach to the development of AI technologies; *“I think that it should be chosen by the community (i.e., the set of people who interact in a meaningful way everyday and share space). Workers, housewives, homeless people, youth. It should be a local, collective decision, not something done by scientists or capitalists or even the government.”* (P54). These comments indicate that the general public is aware of the complexity of deriving moral standards.

Adjacent to the aforementioned concept, a number of participants suggest that the

moral standards should be based on a large-scale democratic outcome among the wider population. “*Sounds like the only fair way to do it is have a referendum of all citizens and go with the popular vote.*” (P09). However, these participants were typically quick to point out the practical limitations of such an approach; “*These need to be based on widely debated and agreed moral codes. Though the practicality of such a process means it would be unlikely to be borne out in practice.*” (P45). Finally, participants also note that even if standards are agreed on by a large sample of a country’s population, they may not necessarily fit the moral standards of other populations; “*I am British and believe we should have a common code or set of values that people should adhere to. E.g. if it is a ‘British Value’ to not discriminate against someone because of their sex or sexuality etc., then mobile systems ought to be able to include those types of moral values. If someone does not agree with that, then they do not adhere to British Values and either have to accept it or find somewhere that does accept that type of discrimination.*” (P62).

5.2. Quantitative analysis

First, we calculate the participants’ level of immersion between the two stimuli conditions (*i.e.*, non-baseline conditions). Transportation Scale [3] results are similar between conditions, with a mean score of 4.7 and 4.8 for the Fable - Positive and Fable - Negative condition respectively. These scores are a bit higher than average immersion scores reported in [3], but given their high similarity are a good indicator of the comparability between the two stimuli.

Second, we calculate the participant’s moral score as based on the completed Moral Identity Questionnaire [10] (MIQ). MIQ scores range from a minimum score of 20 to a maximum score of 120. Average score among our participants is 98.5 (SD = 12.5), with minimal differences between conditions; 98.4, 95.9, and 99.7 for the Fable - Positive, Fable - Negative, and Control condition respectively. In comparing the participant’s self-reported moral identity scores, we find higher average female scores (female mean = 99.2, SD = 11.6 – male mean = 96.1, SD = 11.4), as shown in Figure 2-A. A t-test indicates no significant difference between the two conditions ($t(57) = 2.254$, $p = 0.33$), although the direction in gender difference is in line with earlier work [10]. Following, we analyze the effect of age on the participants’ moral identity. Using a Pearson’s product-moment correlation test, we find a positive but non-significant correlation between participant’s age and participant’s moral identity, $r(57) = 1.45$, $p = 0.15$ (see Figure 2-B). Analysing the Moral Self sub-scale of the MIQ [10], we find a significant correlation ($r = 0.29$) between participant age and moral self score ($r(57) = 2.25$, $p = 0.03$).

Third, we report whether participants’ believe their context affects their moral values. Participants were quite evenly split on this question. A total of 26 participants believe that context would not effect their morality, 29 participants thought context did have an effect on their moral values. None of the participants answered as unsure.

6. Discussion

The call for moral guidelines surrounding digital agents is not new, as was already stated in the seminal work by Rosalind Picard: “*The greater the freedom of a machine, the more it will need moral standards*” [72]. With AI’s increased ubiquity, both in terms of adoption and its concealed nature, this call for moral standards has become increasingly

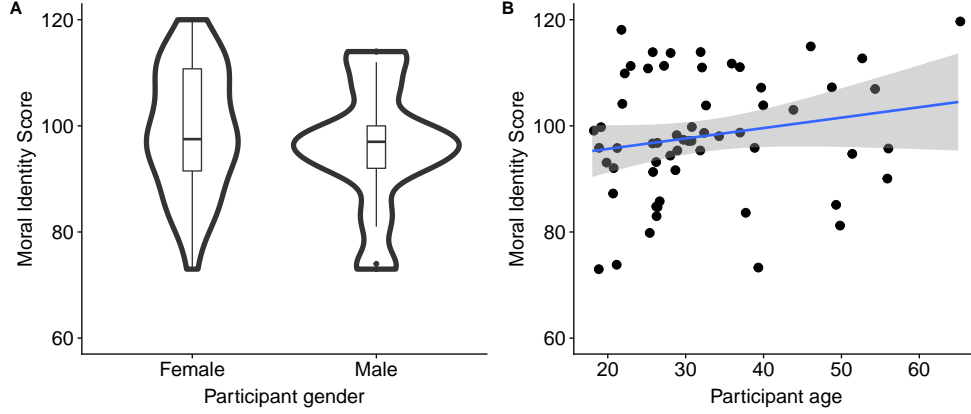


Figure 2. Distribution of participant responses across participant’s (self-identified) gender and age.

pressing. Our qualitative results stress that morality is not a static concept, and that personal AI-systems therefore face unique challenges which have remained underexposed in the current AI-research landscape. The implications of this perspective touch upon multiple aspects of AI research. In this section, we outline how our results build upon previous work and offer an outline for future work in the area of contextual morality.

6.1. *Capturing insights on morality*

In this study we relied on the narrative transportation method to (temporarily) alter the attitudes and intentions of our participants [67, 38, 40, 54]. Our results indicate only minor differences between conditions on the participant’s self-identified moral identity score, which raises interesting points of discussion for future work in this domain. The fables are a clearly moral-infused primer, potentially resulting in participants experiencing resistance to the topic of the study. Literature in Psychology has labeled this behaviour as ‘reverse priming’ [34], stating that participants may attempt to ‘over-respond’ in their choices – consequently biasing their choices in the reverse direction. An alternative explanation can be found in the ‘transformative choices’ theory by Chang [16], which states that moral transformations happen either as a consequence of events (event-based) after we have made a choice, or as a follow-up event of a choice (choice-based). Based on this reasoning, we can argue that our study setup was solely able to assess event-based moral transformations (*i.e.*, making the choice) as opposed to the potential moral transformation that an individual may experience as the consequences of choices they make. A long-term evaluation would be required to assess these choice-based moral transformations.

Previous work included highly engaging stimuli, such as war simulation games [40], resulting in an increase in people’s moral sensitivity. Based on the combined results of previous work and the limited effect our chosen stimuli on participant morality in comparison with the baseline condition, we conclude that stimuli need to be sufficiently stimulating for participants to experience narrative transportation and subsequently experience a change in their morality. Patil *et al.* show how Virtual Reality (VR) could be used as a potential research tool for this purpose [71]. In their experiments,

participants faced variations of the ‘trolley problem’⁵. Participants were found to take a more utilitarian approach (*i.e.*, act proactively to minimize the number of casualties) when shown the same dilemma in VR as compared to a text-based control condition. Through the use of VR simulations, participants can be easily transported into a controlled environment with dynamic stimuli, increasing the ecological validity of the study [69]. In addition to the use of the hypothetical trolley problem scenario, a promising alternative is to derive moral information based on real-life decisions. Game theory provides multiple scenarios, *e.g.* the prisoner’s dilemma, which can be adjusted to mimic realistic decision-making events encountered in daily life.

Finally, an approach which has thus far not been considered in this domain is the collection of participant data ‘in the wild’. An approach particularly suitable for the collection of data in particular contexts (*e.g.*, nightlife areas, at public events) is the use of public displays. Public displays are capable of “*collecting data serendipitously and largely on autopilot — for purposes beyond one single application at a time*” [47], which makes it suitable to collect highly contextual information from a large crowd with a relatively straightforward setup. Although public displays are suitable to collect highly targeted contextual data, they are not well suited to obtain information on the effect of contextual changes at an individual level. With the use of participants’ personal smartphones, the effect of contextual changes can be captured more easily. By completing tasks or answering short questionnaires throughout the day, a method known as Experience Sampling [81], researchers can obtain a rich insight into the life of individuals. These human contributions can be augmented with sensor data collected continuously from the participant’s smartphone, linking the participant’s contributions to specific contextual settings and cognitive states [79]).

6.2. Applying the Contextual Morality Framework to Crowdsourced Data

Despite the proven use of online crowdsourced data in a wide range of application areas, previous work has highlighted the overrepresentation of US-based participants [68], the relative young age [9], and the liberal-leaning ideology [9] of crowdworkers on popular platforms such as Mechanical Turk. Although these samples are still more representative to the US population than in-person convenience samples (*e.g.*, students) on a majority of variables [9], this skewness of the participant sample is likely to introduce biases when collecting morally sensitive questions. This skewness becomes considerably worse when considering the fact that the algorithms based on the contributions from primarily US-based crowdworkers are applied to a worldwide audience. As identified by Shankar *et al.*, the current lack of geo-diversity in (open) data sets results in Amerocentric and Eurocentric representation bias [78]. For example, Shankar *et al.* show that photos of bridegrooms from Ethiopia and Pakistan are not classified as consistently as photos of bridegrooms from Australia and the United States. This is the result of training datasets containing predominantly brides dressed in a white dress as opposed to, *e.g.*, a wedding sari.

In order to enable a human-centred AI perspective, it is key to enrich crowdsourced data with contextual information. Awad *et al.* highlight cultural differences in participants’ viewpoints on ethical dilemmas for autonomous driving [7]. Based on participants’ social expectations, the authors are able to identify three major clusters (‘Western’,

⁵The trolley problem is an ethical thought experiment. In its traditional form, a participant is asked to choose to either intervene in the trajectory of an oncoming trolley and save the live of five persons at the cost of killing one person – or to refrain from action and let the trolley kill five innocent persons.

‘Eastern’, and ‘Southern’), with significant differences identified between the clusters. For example, the preference to save younger people as opposed to older people is strongest in the Southern cluster, and much less pronounced in the Eastern cluster. Similarly, the Southern cluster puts a higher emphasis of saving those with a higher status as compared to the Eastern and Western cluster, as well as a stronger emphasis on sparing females. Furthermore, there are also within-cluster differences – leading to further clustering as for example a Scandinavian and a Commonwealth cluster within the Western cluster. These examples point to the importance of diverse representation in data collection for applications deployed at global scale.

In addition to the aforementioned demographic and cultural disconnect, we also highlight the more general disconnect between those who (indirectly) influence future algorithms, (*i.e.*, crowdworkers who label data) and those who will be affected by the algorithm. This concern, as clearly identified in our qualitative results, can be addressed by developing algorithms in close cooperation with those who are either affected by or actively involved in using the prospective technology. An especially noteworthy example of this perspective is the work by Woodruff *et al.* [86]. Through a series of workshops and interviews with participants from traditionally marginalized populations, the researchers uncover what is of critical importance to the community. Although connecting with those likely to be affected by future algorithms will require additional effort from researchers and developers, this user-centred design approach is critical in aligning expectations and obtaining perspective from relevant stakeholders.

6.3. *Practical Implications of Morality in AI*

Although the tasks completed by a selection of AI applications are sufficiently straightforward to avoid moral scrutiny, user-facing algorithms typically make far-reaching decisions or suggestions. Here, we outline three practical and real-life examples to indicate the possible implications of an AIs (mis)judgement of contextual morality for end users. These examples highlight how context can significantly impact users moral viewpoint as well as their expectations with regards to AI applications.

6.3.1. *Phrase suggestion*

The suggestion of individual or combined words (*i.e.*, phrases) while typing text is an example of a widespread, user-facing, intelligent suggestion algorithm. These suggestion systems, typically personalized based on the user’s past writing style, aim to increase the user’s writing speed. Recent product deployments, such as GMail’s ‘Smart Reply’⁶, which suggests three quick responses based on the content of an email and past writing behaviour, extend this concept by allowing the user to immediately write and send a short message at the tap of a button. As communication with our peers is by definition personal, text-suggestion systems run the risk of disturbing a delicate balance between technology and human-to-human interaction. Recent work by Arnold *et al.* argues that an accurate prediction (*i.e.*, a suggestion which the user would want to select during typing) does not necessarily make for a valuable suggestion [5]. In addition, phrase suggestion raises ethical questions surrounding authorship, freedom of expression, and originality [5] – Arnold *et al.* ask “*if the system tends to offer suggestions with positive valence, will that bias the user towards writing a more positive review?*” [5]. Given these

⁶Google AI blog feature announcement at <https://ai.googleblog.com/2017/05/efficient-smart-reply-now-for-gmail.html>

aforementioned challenges, we argue that phrase suggestion systems would benefit from a better understanding of the user’s contextual morality. Rather than solely considering message content and the user’s historic writing behaviour, elements such as, *inter alia*, the relationship between the user and recipient, time of day, and the user’s mental state can all considerably affect the appropriateness of the suggested phrases.

6.3.2. Facial expression recognition

In our day-to-day interactions, facial expressions can (at least partially) inform us of the emotional state of our conversation partners. Although it is widely agreed that there is an association between our facial expressions and our current emotional state, this association is often times loose and varies between cultures [62, 75]. Facial expression recognition is increasingly used in consumer-facing technology, for example in entertainment [64] and augmented communication applications [77]. However, as these technologies are deployed on a global scale they can quickly run into problems in mapping our emotional state to our moral intuition. As noted by Russell, even the categorization of emotions currently upheld in the English-speaking world are unlikely to map directly to other languages and cultures “*We speakers of English find it plausible that our concepts of anger, fear, contempt, and the like are universal categories, exposing nature at the joints. One way to overcome the influence of such implicit assumptions is to emphasize alternative conceptualizations.*” [75].

6.3.3. Content recommendation and moderation

Online content platforms such as YouTube, Netflix, and Google News strongly rely on continuous content suggestions to offer new and relevant content to their users, thereby sustaining engagement and increasing product usage. These suggestions are based on two main sources of information, namely user data (*e.g.*, ratings, past viewing behaviour) and content meta data (*e.g.*, item titles) [21]. However, as noted by our participants, expected behaviour (both of individuals and of their accompanying technology) can shift depending on context. For example, displays of nudity in TV shows may be undesired when in the presence of guests or family members, and notifications on what is new in show business may be deemed inappropriate in a work setting. Secondly, content moderation – a distressing task often completed through manual annotation – is often complicated by the wide range of moral viewpoints that can be applied. This tension is naturally extended when moderators are not accustomed to the ‘local’ moral standards, potentially overseeing or misjudging defamatory content; “*When CCM [Commercial Content Moderation] work is outsourced to other parts of the world, it creates an additional challenge, in that those workers must become steeped in the racist, homophobic, and misogynist tropes and language of another culture*” [74].

6.4. Contextual Morality in Other Ethical Frameworks

In this paper we focus primarily on a utilitarian perspective in relation to contextual morality, in which the most ethical decision is the one that maximizes value (*i.e.*, utility) [65]. This approach has seen uptake in both the AI and HCI community as it allows researchers to quantify the values of a community by presenting alternative outcome options, *e.g.* the expected behaviour of autonomous cars in calamity situations [11, 7]. We subsequently discuss the presented Contextual Morality Framework in relation to virtue ethics and deontological ethics, two distinctive and dominant families of moral

philosophy in addition to utilitarianism.

Virtue ethics, shortly introduced in the Related Work section, describes the belief that the most moral action is the action taken by a *virtuous* person. Virtue ethics can be traced back to Aristotle [4], who argued that a virtuous person exhibits certain character traits (*e.g.*, courage, modesty). Although it may appear that a trait-based perspective is constant over long periods of times, Govindarajulu et al., in specifying the semantics of a virtuous machine, introduce ‘fluents’ to represent the state of the world in which a decision needs to be made. These type of systems, in which the meaning and value of terms is context dependent, are called intensional systems [35]. The behaviour of intensional systems is therefore context dependent. By measuring the context of the user and their subsequent behaviour across previously unexplored fluents, the system collects augmented training data which can be used to further refine the behaviour of virtue ethics-based systems. Govindarajulu et al. recently discuss the engineering of virtuous machines [36]. The authors argue that “*if the conditions of stability, consistency, explanatory power, and predictive power hold, then virtuous agents or robots might be easier for humans to understand and interact with (compared to consequentialist or deontological agents or robots)*” [36]. A utilitarian decision (grouped under the consequentialist family) may change following only a minor change in the value calculation. For example, in the autonomous car scenario the age of the pedestrian crossing the street (17 versus 18 years old) may significantly alter the car’s decision path if a different value is assigned to adults versus minors – which may surprise end-users of utilitarian-based AI systems.

Deontological ethics argues that the actions of an individual rather than the consequences of this action are to be assessed as moral or immoral. According to Brodhead et al., “*deontologists primarily define what is ‘good’ or ‘right’ as a function of behavior and the context in which that behavior occurs.*” [12]. One of the commonly presented critiques from deontologists towards utilitarianism is that it is oftentimes impractical to consider all potential consequences of all potential actions in order to arrive at the morally right option. This is considered as impractical and inefficient both in terms of energy and time. Such considerations may also affect AI-powered systems, operating under significant time-pressure and continuously updating information on its surroundings. Returning to the autonomous driving example, work by Awad et al. revealed that certain cultures may value the life of *e.g.* a successful business person as more worthwhile. It is unlikely that such information is immediately available when an AI is faced with a situation in which a pedestrian suddenly appears in front of a car.

Our Contextual Morality Framework proposes a continuously evolving interaction between AI development and user context to inform the behaviour of the AI system (Figure 1). Both virtue and deontological ethics indeed require contextual information to arrive at a decision as to whether behaviour is ethical. We therefore conclude that the iterative approach on which our Contextual Morality Framework is based can be beneficial to inform AI-powered systems based on all three discussed ethical families of moral philosophy. Although we argue that the decision-making process of a utilitarian-based AI can be more easily visualised and updated with human input due to the value-based reasoning that underpins the decision-making process, we encourage the interested reader to also consider alternative moral underpinnings in their future research.

6.5. Limitations and future work

The work presented in this paper provides an initial exploration on the importance of contextual morality for AI. As such, we outline multiple avenues for future work. In particular, we believe virtual reality and *in situ* studies offer an unexplored opportunity to measure the effect of context on people’s morality and moral expectations. Furthermore, although our work highlights some of the issues which can be encountered when working with an online crowdsourcing deployment – it provides an opportunity for future work to explore how to mitigate some of the identified issues. Given the explorative nature of this work, it is important to consider a number of (deliberate) limitations. First, our sample of 62 participants consists primarily of Western crowdworkers. As moral differences between cultures are already more widely explored in the literature, we instead focus on the effect of everyday contexts and therefore aimed to control for the participant’s culture by recruiting solely Western participants. Second, the stimuli presented in this work, *i.e.*, the stories presented in the two ‘fable’ conditions, can be considered as relatively unstimulating in comparison to, *e.g.*, related work on the effect of playing as a character in a war-simulation video game [40]. Surprisingly, our participants report relatively high immersion scores as compared with existing work in the literature [3]. Third, we were limited in the ways in which we obtained measurements of the participants’ morality. Future work should consider the use of tests with a tangible and quantifiable outcome (*e.g.*, Prisoner’s Dilemma as used in Economics and Psychology) in addition to the use of self-report questionnaires.

7. Conclusion

This paper investigates the implications of contextual morality for Artificial Intelligence applications. We reveal a number of perspectives on determining suitable moral standards, as outlined by our participant pool. Through an online study employing textual narrative transportation, we find no proof of the effect of the chosen narration on participant’s self-reported morality score. We describe potential avenues for future work into this underexplored domain, in which (virtual) simulation or ‘in the wild’ studies are likely to play an important role. Finally, we argue how morality and FAT* affect each other, and that one cannot be seen in separation from the other – calling for a larger emphasis on context through ‘FACT’. We are hopeful that an increased focus on contextual morality in AI research and application development will result in a more human-centred AI for all users.

8. Acknowledgements

This project was funded by GenZ, a strategic profiling project supported by the Academy of Finland (project number 318930) and the University of Oulu.

References

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference*

- on *Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. DOI:<http://dx.doi.org/10.1145/3173574.3174156>
- [2] Aws Albarghouthi and Samuel Vinitzky. 2019. Fairness-Aware Programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 211–219. DOI:<http://dx.doi.org/10.1145/3287560.3287588>
 - [3] Markus Appel, Timo Gnambs, Tobias Richter, and Melanie C. Green. 2015. The Transportation Scale–Short Form (TS–SF). *Media Psychology* 18, 2 (2015), 243–266. DOI:<http://dx.doi.org/10.1080/15213269.2014.987400>
 - [4] Aristotle. 2000. *Aristotle: Nicomachean Ethics*. Cambridge University Press. DOI:<http://dx.doi.org/10.1017/CB09780511802058>
 - [5] Kenneth C. Arnold, Kai-Wei Chang, and Adam Tauman Kalai. 2017. Counterfactual Language Model Adaptation for Suggesting Phrases. *CoRR* abs/1710.01799 (2017).
 - [6] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2018. FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity. *CoRR* abs/1808.0 (2018).
 - [7] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (2018), 59. DOI:<http://dx.doi.org/10.1038/s41586-018-0637-6>
 - [8] Solon Barocas and Andrew D Selbst. 2016. Big data’s disparate impact. *California Law Review* 104 (2016), 671.
 - [9] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368. DOI:<http://dx.doi.org/10.1093/pan/mpr057>
 - [10] Jessica E. Black and William M. Reynolds. 2016. Development, reliability, and validity of the Moral Identity Questionnaire. *Personality and Individual Differences* 97 (2016), 120–129. DOI:<http://dx.doi.org/10.1016/j.paid.2016.03.041>
 - [11] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576. DOI:<http://dx.doi.org/10.1126/science.aaf2654>
 - [12] Matthew T. Brodhead, David J. Cox, and Shawn P. Quigley. 2018. Chapter 1 - Introduction to ABA, Ethics, and Core Ethical Principles. In *Practical Ethics for Effective Treatment of Autism Spectrum Disorder*, Matthew T. Brodhead, David J. Cox, and Shawn P. Quigley (Eds.). Academic Press, 1 – 16. DOI:<http://dx.doi.org/10.1016/B978-0-12-814098-7.00001-8>
 - [13] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 4, 14 pages. DOI:<http://dx.doi.org/10.1145/3290605.3300234>
 - [14] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (1 Sept. 2010), 277–292. DOI:<http://dx.doi.org/10.1007/s10618-010-0190-x>
 - [15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie

- Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1721–1730. DOI:<http://dx.doi.org/10.1145/2783258.2788613>
- [16] Ruth Chang. 2015. Transformative Choices. *Res Philosophica* 92, 2 (2015), 237–282. DOI:<http://dx.doi.org/10.11612/resphil.2015.92.2.14>
- [17] Amit K Chopra and Munindar P SIngh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 48–53. DOI:<http://dx.doi.org/10.1145/3278721.3278740>
- [18] Kate Crawford. 2016. Artificial intelligence’s white guy problem. (2016).
- [19] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature News* 538, 7625 (2016), 311. DOI:<http://dx.doi.org/10.1038/538311a>
- [20] Samuel Croxall and others. 1775. *Fables of Æsop and Others*. W. Strahan.
- [21] James Davidson, Benjamin Liebald, Junjing Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. ACM, New York, NY, USA, 293–296. DOI:<http://dx.doi.org/10.1145/1864708.1864770>
- [22] Munmun De Choudhury and Emre Kiciman. 2018. Integrating Artificial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health. *AI Magazine* 39, 3 (2018), 69–80.
- [23] Anind K. Dey. 2001. Understanding and Using Context. *Personal Ubiquitous Comput.* 5, 1 (Jan. 2001), 4–7. DOI:<http://dx.doi.org/10.1007/s007790170019>
- [24] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun.* ACM 59, 2 (2016), 56–62.
- [25] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018). DOI:<http://dx.doi.org/10.1126/sciadv.aao5580>
- [26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226. DOI:<http://dx.doi.org/10.1145/2090236.2090255>
- [27] Jacques Ellul. 1964. *THE TECHNOLOGICAL SOCIETY*. New York: Knopf.
- [28] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 153–162. DOI:<http://dx.doi.org/10.1145/2702123.2702556>
- [29] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes Towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 494, 14 pages. DOI:<http://dx.doi.org/10.1145/3290605.3300724>
- [30] Europost. 2019. G20 ministers agree on human-centric AI principles. (June 2019).
- [31] Beverley Garrigan, Anna L.R. Adlam, and Peter E. Langdon. 2018. Moral decision-making and moral development: Toward an integrative framework. *Developmental*

- Review* 49, June (2018), 80–100. DOI:<http://dx.doi.org/10.1016/j.dr.2018.06.001>
- [32] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Dauméé, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* (2018), 1–27.
 - [33] Bernard Gert and Joshua Gert. 2017. The Definition of Morality. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
 - [34] Jack Glaser and Mahzarin R Banaji. 1999. When fair is foul and foul is fair: reverse priming in automatic evaluation. *Journal of Personality and Social Psychology* 77, 4 (1999), 669. DOI:<http://dx.doi.org/10.1037/0022-3514.77.4.669>
 - [35] Naveen Sundar Govindarajulu and Selmer Bringsjord. 2017. On Automating the Doctrine of Double Effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, 4722–4730.
 - [36] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the Engineering of Virtuous Machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 29–35. DOI:<http://dx.doi.org/10.1145/3306618.3314256>
 - [37] Melanie C. Green and Timothy C. Brock. 2002. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, Chapter In the mind’s eye: Transportation-imagery model of narrative persuasion, 315–341.
 - [38] Melanie C. Green and Jenna L. Clark. 2013. Transportation into narrative worlds: implications for entertainment media influences on tobacco use. *Addiction* 108, 3 (2013), 477–484. DOI:<http://dx.doi.org/10.1111/j.1360-0443.2012.04088.x>
 - [39] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74, 6 (1998), 1464–1480.
 - [40] Matthew Grizzard, Ron Tamborini, Robert J. Lewis, Lu Wang, and Sujay Prabhu. 2014. Being Bad in a Video Game Can Make Us Morally Sensitive. *Cyberpsychology, Behavior, and Social Networking* 17, 8 (2014), 499–504. DOI:<http://dx.doi.org/10.1089/cyber.2013.0658> PMID: 24950172.
 - [41] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 3323–3331.
 - [42] Gilbert Harman. 1999. Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society* 99 (1999), 315–331.
 - [43] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29.
 - [44] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts?: The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 415, 11 pages. DOI:<http://dx.doi.org/10.1145/3290605.3300645>
 - [45] Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture* 2, 1 (2011), 8. DOI:<http://dx.doi.org/10.9707/2307-0919.1014>
 - [46] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and

- Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 600, 16 pages. DOI:<http://dx.doi.org/10.1145/3290605.3300830>
- [47] Simo Hosio, Andy Alorwu, Niels van Berkel, Miguel Bordallo López, Mahalakshmy Seetharaman, Jonas Oppenlaender, and Jorge Goncalves. 2019. Fueling AI with Public Displays?: A Feasibility Study of Collecting Biometrically Tagged Consensual Data on a University Campus. In *Proceedings of the 8th ACM International Symposium on Pervasive Displays (PerDis '19)*. ACM, New York, NY, USA, Article 14, 7 pages. DOI:<http://dx.doi.org/10.1145/3321335.3324943>
- [48] Don Howard and Ioan Muntean. 2017. *Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency*. Springer International Publishing, Cham, 121–159. DOI:http://dx.doi.org/10.1007/978-3-319-61043-6_7
- [49] Ellora Thadaneey Israni. 2017. When an Algorithm Helps Send You to Prison. (2017).
- [50] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Kattia Sycara. 2018. Transparency and Explanation in Deep Reinforcement Learning Neural Networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 144–150. DOI:<http://dx.doi.org/10.1145/3278721.3278776>
- [51] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 421, 12 pages. DOI:<http://dx.doi.org/10.1145/3173574.3173995>
- [52] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. DOI:<http://dx.doi.org/10.1038/s42256-019-0088-2>
- [53] Dan R. Johnson, Brandie L. Huffman, and Danny M. Jasper. 2014. Changing Race Boundary Perception by Reading Narrative Fiction. *Basic and Applied Social Psychology* 36, 1 (2014), 83–90. DOI:<http://dx.doi.org/10.1080/01973533.2013.856791>
- [54] Dan R Johnson, Daniel M Jasper, Sallie Griffin, and Brandie L Huffman. 2013. Reading Narrative Fiction Reduces Arab-Muslim Prejudice and Offers a Safe Haven From Intergroup Anxiety. *Social Cognition* 31, 5 (2013), 578–598.
- [55] Rachana Kamtekar. 2004. Situationism and Virtue Ethics on the Content of Our Character. *Ethics* 114, 3 (2004), 458–491. DOI:<http://dx.doi.org/10.1086/381696>
- [56] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review* 165 (2016), 633.
- [57] Olya Kudina and Peter-Paul Verbeek. 2019. Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy. *Science, Technology, & Human Values* 44, 2 (2019), 291–314. DOI:<http://dx.doi.org/10.1177/0162243918793711>
- [58] Bruno Latour and Couze Venn. 2002. Morality and Technology. *Theory, Culture & Society* 19, 5-6 (2002), 247–260. DOI:<http://dx.doi.org/10.1177/026327602761899246>
- [59] Keith Leavitt, Scott J. Reynolds, Christopher M. Barnes, Pauline Schilpzand, and

- Sean T. Hannah. 2012. Different Hats, Different Obligations: Plural Occupational Identities and Situated Moral Judgments. *Academy of Management Journal* 55, 6 (2012), 1316–1333. DOI:<http://dx.doi.org/10.5465/amj.2010.1023>
- [60] Susan Leavy. 2018. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (GE '18)*. ACM, New York, NY, USA, 14–16. DOI:<http://dx.doi.org/10.1145/3195570.3195580>
- [61] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (1 Dec. 2018), 611–627. DOI:<http://dx.doi.org/10.1007/s13347-017-0279-x>
- [62] Kristen A Lindquist, Lisa Feldman Barrett, Eliza Bliss-Moreau, and James A Russell. 2006. Language and the perception of emotion. *Emotion* 6, 1 (2006), 125.
- [63] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 3150–3158.
- [64] Adam Lobel, Marientina Gotsis, Erin Reynolds, Michael Annetta, Rutger C.M.E. Engels, and Isabela Granic. 2016. Designing and Utilizing Biofeedback Games for Emotion Regulation: The Case of Nevermind. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1945–1951. DOI:<http://dx.doi.org/10.1145/2851581.2892521>
- [65] J.S. Mill. 1863. *Utilitarianism*. Parker.
- [66] Susan E. Morgan, Lauren Movius, and Michael J. Cody. 2009. The Power of Narratives: The Effect of Entertainment Television Organ Donation Storylines on the Attitudes, Knowledge, and Behaviors of Donors and Nondonors. *Journal of Communication* 59, 1 (2009), 135–151. DOI:<http://dx.doi.org/10.1111/j.1460-2466.2008.01408.x>
- [67] M. Morgenstern, E. A. Poelen, R. Scholte, S. Karlsdottir, S. H. Jonsson, F. Mathis, F. Faggiano, E. Florek, H. Sweeting, K. Hunt, J. D. Sargent, and R. Hanewinkel. 2011. Smoking in movies and adolescent smoking: cross-cultural study in six European countries. *Thorax* 66, 10 (Oct. 2011), 875–883.
- [68] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making* 5, 5 (2010), 411–419.
- [69] Thomas D. Parsons. 2015. Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences. *Frontiers in Human Neuroscience* 9 (2015). DOI:<http://dx.doi.org/10.3389/fnhum.2015.00660>
- [70] Frank Pasquale. 2015. *The Black Box Society: the Secret Algorithms that Control Money and Information*. Harvard University Press.
- [71] Indrajeet Patil, Carlotta Cogoni, Nicola Zangrando, Luca Chittaro, and Giorgia Silani. 2014. Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience* 9, 1 (2014), 94–107. DOI:<http://dx.doi.org/10.1080/17470919.2013.870091>
- [72] Rosalind W Picard. 1997. *Affective computing*. MIT press.
- [73] Scott J Reynolds, Keith Leavitt, and Katherine A DeCelles. 2010. Automatic ethics: the effects of implicit assumptions and contextual cues on moral behavior.

- The Journal of Applied Psychology* 95, 4 (July 2010), 752—760. DOI:<http://dx.doi.org/10.1037/a0019411>
- [74] Sarah Roberts. 2016. *Commercial Content Moderation: Digital Laborers’ Dirty Work*.
 - [75] James A Russell. 1994. Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies. *Psychological Bulletin* 115, 1 (1994), 102.
 - [76] Julian Savulescu and Hannah Maslen. 2015. *Moral Enhancement and Artificial Intelligence: Moral AI?* Springer International Publishing, Cham, 79–95. DOI: http://dx.doi.org/10.1007/978-3-319-09668-1_6
 - [77] S. A. Scherr, F. Elberzhager, and K. Holl. 2018. Acceptance Testing of Mobile Applications - Automated Emotion Tracking for Large User Groups. In *2018 IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. 247–251.
 - [78] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NIPS 2017 workshop: Machine Learning for the Developing World*.
 - [79] Benjamin Tag, Andrew W. Vargo, Aman Gupta, George Chernyshov, Kai Kunze, and Tilman Dingler. 2019. Continuous Alertness Assessments: Using EOG Glasses to Unobtrusively Monitor Fatigue Levels In-The-Wild. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 464, 12 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300694>
 - [80] Upstill Trystan. 2018. The new Google News: AI meets human intelligence. (May 2018).
 - [81] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6, Article 93 (Dec. 2017), 40 pages. DOI:<http://dx.doi.org/10.1145/3123988>
 - [82] Niels van Berkel, Simo Hosio, Benjamin Tag, and Jorge Goncalves. 2019. Capturing Contextual Morality: Applying Game Theory on Smartphones. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. DOI:<http://dx.doi.org/10.1145/3341162.3344846>
 - [83] Tom van Laer, Ko de Ruyter, Luca M. Visconti, and Martin Wetzels. 2013. The Extended Transportation-Imagery Model: A Meta-Analysis of the Antecedents and Consequences of Consumers’ Narrative Transportation. *Journal of Consumer Research* 40, 5 (Aug. 2013), 797–817. DOI:<http://dx.doi.org/10.1086/673383>
 - [84] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, New York, NY, USA, Article 440, 14 pages. DOI: <http://dx.doi.org/10.1145/3173574.3174014>
 - [85] Scott J. Vitell, Saviour L. Nwachukwu, and James H. Barnes. 1993. The effects of culture on ethical decision-making: An application of Hofstede’s typology. *Journal of Business Ethics* 12, 10 (1993), 753–760. DOI:<http://dx.doi.org/10.1007/BF00881307>
 - [86] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, New York, NY, USA, Article 656, 14 pages. DOI:

- <http://dx.doi.org/10.1145/3173574.3174230>
- [87] David C Wyld and Coy A Jones. 1997. The Importance of Context: The Ethical Work Climate Construct and Models of Ethical Decision Making – An Agenda for Research. *Journal of Business Ethics* 16, 4 (March 1997), 465–472. DOI:<http://dx.doi.org/10.1023/A:1017980515603>
 - [88] Daniel Yudkin, Ana Gantman, Wilhelm Hofmann, and Jordi Quoidbach. 2019. Moral Values Gain Importance in the Presence of Others. (2019). DOI:<http://dx.doi.org/10.31234/osf.io/tcq65>
 - [89] Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2012. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. *Artificial Intelligence* 184-185 (2012), 17–37. DOI:<http://dx.doi.org/10.1016/j.artint.2012.02.002>
 - [90] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it’s time to make it fair. 559 (2018), 324–326. DOI:<http://dx.doi.org/10.1038/d41586-018-05707-8>