Foundational Questions About Values in Information Technology

Fiorella Battaglia *
fiorella.battaglia@unisalento.it

ABSTRACT

In the contemporary debate about values, information technology constitutes an important source of hard ethical questions and in turn is a testing area for the moral theory of values. Values are difficult to track down and yet there are a number of inquiries starting from economics, social psychology, ethics, and political theory that engage with the cognitive, epistemic, and moral status of values. This paper is a contribution to an account of values in connection with information technology. It argues that information technology may provide further support to a theory of values that is able to embrace the transformative effects of the digital revolution. In particular, it is plausible that a non-ideal reflection on digital wrongdoings is better equipped to produce substantive knowledge about values that have been undermined than a different approach focused on ideal guiding values. Moreover, information technology overcomes the vaunted fact/value dichotomy and supports the fact/value entanglement. As the principal concern of data-mining and machine-learning communities are ways of remedying a remarkable number of biases and conformism in techno-social systems, it is within the bounds of possibility to supplement the non-ideal theory from this new practical angle. I therefore call for a fully conceptual consideration of values drawing on the experience and reflection that is growing in the field of information technology.

1. Introduction

What, then, are values? If no one asks us, we know; if we wish to explain it to the enquirer, we do not know. To answer this hard question about what is of some value to us, various disciplines have made their contributions (Brosch

* Università del Salento, Lecce, Italy, and Ludwig-Maximilians-Universität, München, Germany.

and Sander 2015). In particular, moral values are a hotly debated topic in philosophy. According to Hilary Putnam, the source of this question has to be retrieved in the distinction about 'crime' made by Hume between fact and value judgment, which 'arises from a complication of circumstances that when presented to the spectator excites the *sentiment* of blame, by the particular structure and fabric of his mind' (Hume 1978). By placing values in our minds and thus removing them from the group of facts that serve to describe and explain the physical world, Hume prepares the ground for making values something subjective as opposed to the objectivity of natural facts. The subjective character has been interpreted as something thoroughly 'up to us', thus bordering on the arbitrary. However, the relational character that binds values to the subject need not be interpreted in this way. The genesis of the fact/value dichotomy has two other remarkable steps: the Kantian philosophy and its elaboration in logical empiricism. Attention to values has also been paid outside the discipline of philosophy, as values are an elusive and yet ubiquitous phenomenon of everyday practices. Interdisciplinary work has therefore shaped the debate on values. What is certain is that these various perspectives share a practical approach. Each disciplinary field contributes to staking out various elements in this nexus of values.

2. Making conflicting values safe for democracy

The concept of 'value' acquires its meaning in relation to the economic sphere. Value receives its meaning within the context of the market. In particular, it signifies an evaluation of the 'public good' by the individual who will make use of it. In this context, it is all about individuals who attempt to maximize their strategies. This anthropological view sees the human as a *homo oeconomicus* who always optimizes. Within this framework value is just a utility function. As a function it has objective status whereas the other values are expressions of subjective preferences. This view has been criticized, as it tends to reduce the dynamics of values to just one utility function (Sen 2000). The dawn of the economic conception of value finds its legacy in the vaunted dichotomy between fact and value, i.e. between objective and subjective. This distinction is also used when discussing moral realism in metaethics (Railton 1986; Shafer-Landau 2009).

The discourse on values has a distinct political version. This is not surprising, since politics is about actions and reasons. Its political version is

known as "epistemic political liberalism". This position is particularly interesting for values and information technology, as it is committed to delivering benefits to citizens in terms of efficiency, consistency and accuracy (Sunstein 2022). What is more, Sunstein's support towards properly constructed algorithms, which will be able to better the performance of the administrative state, includes the quality and the truth of the given political arrangements (Battaglia 2021). Indeed they do matter for democracy. Deliberative democracy is concerned both with the truth and with ways of eliminating the unequal treatment of citizens. From the point of view of the discourse on values, it is intriguing that there are values that are equally important to us and yet come into conflict. From the perspective of political theory, the inquiry focuses on the epistemic dimension of democracy. The special authority of science as a source of knowledge comes into tension with the essential dimensions of human wellbeing when good policy is at stake. (Estlund 2008; Elkins and Norris 2012). Democracy, as we understand it, is a fair way of making decisions. However, is this demand for fairness in the process compatible with good resulting arrangements? After all, democracy is not known for its tendency to produce good decisions. If we value democracy for its fairness, then we should be happy with a random procedure, because it should be just as good. Again, it is worth focusing on what happens when information technology solutions are implemented, because it can make a huge difference in the knowledge produced.

The most important results of the interdisciplinary research on values are its practical orientation and its focus on competing values. Some strands, however, suggest a radically different perspective on values. How can we characterize the debate prior to the introduction of the perspective? What is the state of the debate in science and technology of information technology? Does it add another perspective?

3. Values in science and values in technology

This section explores the idea that there is a distinct difference between values in science and values in technology. It argues that this difference relies on their specific scope (McGinn 1991). More specifically, there is a correlation between the amount of epistemic effort and prominent values. The quintessence of science is both to free the epistemic process from the grasp of traditional authority and to set the stage for defense against future intrusions

by demarcation. Its historical origin justifies the process of marking the limits and boundaries of epistemic lines. This process has provided a useful demarcation between epistemic rationality and traditional authorities thus sharply separating science from other spheres, such as state and religion (Jasanoff 2005). As a result, the values of the scientific enterprise are internal to the scientific community. Since they concern the characteristics of knowledge as a whole, they have an additional aesthetic quality. Indeed, the values that are held in high esteem in the scientific community are coherence, simplicity, naturalness, and beauty.

As science has become more technical and action-oriented, two things have happened. Firstly, it has enlarged its scope to encompass society at large, so that its values are no longer directed only at the scientific community, but also include citizens beyond that community. Secondly, the values involved are no longer values that refer to epistemic rationality; on the contrary, they signal how knowledge production is linked to scientific responsibility.

I will focus on this crucial shift from knowledge production to action-oriented epistemic practice in my framing of the question of how values can change depending on the scope of knowledge production. Its relationship to society will have deep and profound implications for science as a discipline, profession and practice. Without an awareness of the practical significance of information technology, any account of moral values in the context of human-AI interaction is no longer plausible. On the other hand, in its eagerness to provide practical suggestions, computer science may lose sight of philosophical issues.

4. How can we characterize the discourse on values in information technology? Turning to applied research

One of the main insights that philosophy can draw from the comparison with the arguments developed in other disciplines concerns the epistemic status of values. It is striking that the expressivist challenge is being defeated in information technology. Expressivists deny that values represent the world as being one way rather than another. According to the emotivism they defend, values merely convey non-cognitive attitudes such as desires, preferences, and pros and cons of some other kind (Schroeder 2007). It is the introduction of the agential perspective, i.e. the perspective of an agent who is about to act in order to change the world for the better, that will weaken the argument against

the cognitivist approach to values. Put another way, information technology no longer places values outside the sphere of debates about rationality. In order to achieve this result, it may be necessary to take a circuitous route through a wide range of activities, including identifying and making sense of the values at stake in specific critical situations. In other words, the study of values in the field of information technology can provide us with many valuable insights that can be used to imagine how commitment to moral and democratic values can be developed. This means that, they are not derived from mere rational calculation, but through a series of critical views and studies based both on exposure to a variety of value conceptions and on confrontation with the results derived from monitoring and analysing automated systems and their impact on society. This is especially true for biases, which jeopardise moral values. Only in this unique way can the ineluctable character of techno-social systems, i.e. to be value-laden and inherently morally motivated, unfold its positive impact on people and society.

The proposed detour has not only practical consequences, but also theoretical ones concerning the cognitive status of values. Expressionism is defeated (Stevenson 1937). This is the benefit of relying on a perspective that aims at operationalisation. One of the main characteristics of technology is that it does not simply aim to improve our understanding of the world, but to understand it in order to change it. Its epistemic concepts are loaded from the start with concerns about avoiding bias, escaping unacceptable generalisations, sorting out non-robust, bias-triggering inferences, and discriminating inferences. They are also concerned with escaping new ontologies that tend to aggregate people according to principles of conformity and, as a result, relegate them to antagonistic and polarised communities (Pariser 2011). To change the world for the better, you have to accept that mistakes can potentially be made. Changing the world for the better also requires the adoption of a new framework, one that is not driven solely by epistemic concerns, but also by agential constraints. It is therefore a kind of platitude to say that science is ethically committed in its technical and economic mediation. Engineers and computer scientists are therefore engaged in activities that are inherently morally motivated. As such, they need ethical literacy that goes beyond their technical skills. Information technology professionals are in a unique position to face ethically complex scenarios that can have a profound impact on the rest of society. Unlike more established professions such as medicine or law, computer scientists and engineers are

beginning to establish shared values (Blanken-Webb et al. 2019). The most important values being discussed in information technology are freedom, justice, fairness and equality. The fact that they are not very different from the moral and political values with which we are already familiar should not come as too much of a surprise to us. What is new, however, is that they can become conflicting, and that we may therefore be forced to weigh one value against another, e.g. privacy versus security. Indeed, there is a gap between adopting new technologies and having equal opportunities to use them. From a more historical perspective, ethical concerns and concepts entered information technology discourses as early as 1980 (Winner). The philosophical examination of information technology and moral values began with the work of the philosopher James Moor, who argued that because information technology gave us new ways of acting, new values would emerge. At the same time, he pointed out that software could embody biases (1985). Empirical work by Chuck Huff and Joel Cooper highlighted the potential for gender bias in educational software (1987). Nissenbaum emphasises two possible perspectives on technology with far-reaching implications for the values discourse. According to Nissenbaum, technology can be understood (i) as given and (ii) as mediation. The second option points to consequences and implications of information technology that are not only technical in nature (1996, 1998). This comprehensive approach is ahead of its time and will pave the way for the argument that ethical, legal and social aspects of information technology should not be seen as an appendix to a classical approach to information technology. Even aspects that do not strictly belong to the technical side of information technology are therefore at the core of the classical approach to information technology. I will argue that while people′s experiences with applications of information technology are in many cases epistemically and ethically unsound, they are nevertheless not practically devoid of effects: they are the materials from which much of the remedying of biases in techno-social systems is designed. It is worth mentioning some theoretical contributions. They are closely related to the phenomenon of bias but avoid its replication. The general idea is that non-ideal constraints on rationality, which are the result of technically informed corrective practices, can provide us with better assumptions in the process of moral reasoning. I am trying to develop an account of moral values that does not require criteria divorced from reality; indeed, my approach focuses on the constraints of feasibility. The success of this development is demonstrated by its ability to

deal with transformative effects. As an intermediate conclusion, we can say that the source of ethical concern is to be found in moral condemnation for reinforcing pre-existing biases and conformism (Susser, Roessler, and Nissenbaum 2019). The exposure of algorithms that discriminate against marginalised populations, and of algorithms that tend to polarise society, works as a trigger for ethical reflection. A practice-based approach characterises the questioning of values in the context of non-ideal theory. In this ′realistic′ account, technical considerations are not abstracted from political, social and economic aspects. In terms of the theoretical framework, it is true that feasibility considerations aimed at overcoming threatened values constrain axiological theorising. The translation of the methodological label ′non-ideal theory′ into the field of information ethics is in need of clarification. It requires explanation because I am arguing for the application of a certain orientation in political philosophy to the field of information ethics. I believe that my argument for an 'extended application' of 'non-ideal theory' to the field of information technology can open up an interesting new line of debate, both for the theory itself and for a number of relevant case studies in information ethics. The Rawlsian original position has been a standard way of trying to determine the nature of justice in the field of political ideal theory. This approach has been criticised. It is incapable of informing real-world policies (Carens 1996; Haslanger 2019). More generally, in recent years much of contemporary normative political theory has been increasingly criticised for abstracting from real and concrete political, social and economic aspects. A kind of methodological debate about the proper nature of political theory and its ability to guide policy in real-world scenarios has become popular as the ′ideal versus non-ideal theory′ debate. What I want to do here is translating this shift from ideal to non-ideal theorising into the field of information ethics. In my view, it is possible to distinguish between two positions in moral theorising. On the one hand, to be more concerned with the question of what issues and judgments should be classified as moral, and on the other, to engage with substantive ethical questions. The latter position will have dramatic implications for the methods and content of moral theory. Thus I will argue.

There are a number of reasons for turning the ideal theory into a non-ideal theory in the field of information technology.

- The *first* reason is of a metaphysical nature: we have to deal with a domain that is by nature practical.

- The *second* reason is that the level of anxiety about techno-social systems threatens to slow down the digitalisation of our world and thus the benefits that could be derived from it. This great concern leads to a preference for forms of argumentation that have the capacity to impact the real world.
- The *third* reason is that since the method used in information technology starts from the identification of anomalies in order to address the remedies much more easily, it seems necessary to formalise this theoretical trend, highlighting the advantages not only in ethical practice but also in moral theory.

In particular, moral theory is more likely to identify and characterise relevant ethical phenomena when confronted with wrongdoing than with ideal characteristics. Contrary to what Augustine thought, evil is not the absence of good, but a phenomenon with a reality of its own. Certainly, analysing the inequalities, stereotypes and conformism in the field of digital networks can help us to identify the remedies. More importantly, it can also help us to develop a more concrete moral theory (Origgi and Ciranna 2017). It can also be used to map the distorted values. It is not possible to arrive at this result by means of a standard a priori method. On the contrary, we must work backwards: from the non-ideal to the ideal. The reason for this is that wrongdoings mediated by technology cannot always be traced back to clear cases of epistemic or ethical failure. It may not be easy to get them to go back to a non-compliance with some of the principles. Undoubtedly, some of the effects may be questionable, but it is hard to give reasons for this on the basis of standard moral theories. The sense of unease that their effects create is based on the fact that we cannot make sense of them by bringing them back to the usual ways of dealing with things (Mittelstadt et al. 2017). A more detailed analysis of the implicit and explicit values at stake is needed in these cases. This analysis should be carried out at different levels and cover different issues. This comprehensive analysis allows us to address other relevant issues, in particular the way in which we conceptualise the world and modify its social and political organisation following the introduction of information technology applications. The mediation of human action and interaction introduced by technology has proved to be a remarkable quality that raises methodological and epistemological questions in the field of artificial morality (Verbeek 2019). I would like to be more explicit about the advantages of the proposed

approach. If we stay within the confines of the standard view, the analysis of ethical issues raised by new technologies is a matter of applying principles and values to the case at hand. Despite its apparent straightforwardness, however, the methodological approach of applied ethics is inherently ambiguous. On the one hand, this kind of ethical work aims to regulate a specific area of human practice. In this sense, it seems very specific and highly contextualised. On the other hand, applied ethics provides a set of principles without challenging them, taking them for granted. They are meant to be part of a general system of moral rules and principles. These moral rules and principles should be applied to specific contexts in order to disentangle right and wrong technological interventions. This a priori type of ethical work is inadequate, not only because it does not question the basis of these principles, but also because of its shortcomings with regard to the novelty and specificity of the new conceptualisation of human practice suggested by technological innovation. Instead, if we let go of the assumption that a ready-made system of values already exists, we will have an ethical theory that integrates new knowledge and its implications into our moral practice. Specific research and application-related problems of moral judgement then become a constitutive part of ethical theory itself. From this perspective, the introduction of constraints of non-ideal rationality in information ethics will produce both working results because of their remarkable practical help and new theoretical insights. It will also have implications for the redress of biases in techno-social systems. From a more theoretical point of view, even if it is not of an ideal nature, this development should be considered as new material to the idea of the entanglement of facts and values. Indeed, from this point of view, the entanglement of facts and values is becoming more significant. For instance, Nissenbaum identified three categories of biases that reflect the complexity of the fact/value entanglement:

> 'We also developed a theoretical framework which identified three categories of bias – namely, preexisting bias (reflecting biases preexisting in society), technical bias (arising form technical constraints), and emergent bias (arising as a result of contextual shift)' (Nissenbaum 1998, 38).

Along these lines, a very recent approach to prejudice and inter-group conflict has shown that there is an emergent kind of bias (Whitaker, Colombo and Rand 2018). They highlight another distinctly epistemic kind of value violation, which is algorithmic in nature. In conclusion, the literature on moral

values and information technology shows that evidence of new insights comes from a variety of experimental protocols. It supports: (i) the methodological commitment to non-ideal theory when it comes to accounting for values in information technology; (ii) it also supports the phenomenon of fact/value entanglement; (iii) finally, it supports working outcomes as a way of remedying bias in techno-social systems.

5. Non-ideal rationality constraints originating in bias and conformism

What approach should be used to try to determine the nature of values? This section explores the epistemology of values in the context of non-ideal theory, with particular attention to the epistemology of redressing bias and conformism in socio-technical systems. While biassed or inaccurate knowledge and the design and management of individuals and groups have given rise to numerous scholarly arguments, their epistemological capital remains unexplored. It may be prudent to work with a bias-derived understanding, since agreement on an overarching definition is weak. According to Nissenbaum, a biassed system is one that systematically and unfairly discriminates against some in favour of others. This is a kind of relational condition since it refers to bias in terms of damaged values. Bias is a particular instance of the more general phenomenon of values being embedded in the design of techno-social systems (Nissenbaum 1998). A bias is not just an error in the cognitive process. Rather, bias is a wrong that is done to someone with respect to a particular value. In other words, bias amounts to a misapplication of some value, i.e., a failure to recognize or appreciate things that we - as individuals and as groups - hold dear. Biases are ways of failing to achieve fair results. Drawing on this relationship between bias and value, I will argue that information technology can develop a 'practice-based' approach to values, whose point of entry is a focus on bias. It is undeniable that digitally transformed practices are part of the social arrangements that are under the influence of a comprehensive view of the good. As with analogue practices, digital practices are also being called into question. When are they morally good, right, or praiseworthy? When are they blameworthy? The goal of social critique applied to algorithmic practices gone wrong is to uncover the nature of

bias as it unfolds in the digital implementation of said practice.[1] One set of legitimate questions in information ethics begins with concerns raised by bias, i.e. by digital wrongdoings. Empirical research suggests that technically feasible systems operating within accepted parameters do not produce ethically sustainable behaviour; very often, biassed behaviour is simply a replication of pre-existing wrongdoing; in other cases, it arises from human-machine interaction. Sally Haslanger′s work on 'ameliorative projects' on race and gender argues for the importance of empirical work over the more speculative task of describing our concepts and their uses:

> 'Just as medicine is done in the service of human health, critical social science is done in the service of justice. Likewise, we learn more about what health is and what justice is by doing value-informed empirical work' (Haslanger 2019, 12).

In an earlier paper, she is more radical in setting the stage for empirical research:

> 'For example, the question 'What is knowledge?' might be construed in several ways. One might be asking: What is our concept of knowledge? ... On a more naturalistic reading one might be asking: What (natural) kind, (if any) does our epistemic vocabulary track? Or one might be undertaking a more revisionary project: What is the point of having a concept of knowledge? What concept, (if any) would do that work best?' (Haslanger 2000).

Her visionary project on conceptual engineering in the social sciences is driven by the same epistemic stance as James Moor and his analysis of the computer revolution. On Moor's account, the permeation stage is characterised by questions like this: 'What is the nature and value of such and such an activity?' (Moor 1985). If we integrate Haslanger′s talk of a social engineering approach and Moor′s talk of a technological revolution, we get an ameliorative project that explores the negative space made available by the dramatic changes

[1] https://algorithmwatch.org/ AlgorithmWatch is a non-profit research and advocacy organisation committed to evaluating and shedding light on algorithmic decision-making processes that are socially relevant, meaning they are used either to predict or prescribe human action or to make decisions automatically.

promised by technology. Social engineering approaches may act in order to empower the citizens and further the common good (values) (Boucher et al. 2018; Cappelen 2018). The first step in this approach is to address the nature of the bias in terms of its novelty or of being the result of a combination of old (analogue) traits with new (digital) tendencies. Nissenbaum has tried to introduce the first classification (1998). She and her group developed a theoretical framework that identified three categories of bias depending on the combination of the technical and social parts of the system. It includes:

1. pre-existing bias (reflecting biases pre-existing in society),
2. technical bias (arising from technical constraints), and
3. emergent bias (arising as a result of contextual shifts).

In more recent times, a new concern has been added. This type of bias arises from the process of personalisation in social networks, which is sometimes autonomously chosen (Sunstein 2008) and sometimes the result of profiling activities (Milano et al. 2020). In both cases, they limit people′s access to viewpoints and issues, and promote conformism and polarisation, which is a rather despicable development for democracy (Bozdag and van den Hoven 2015). Scholars in the field of social epistemology have shown that these new developments in social networks pose a challenge to democracy. They have examined in detail how social learning is affected by conformity bias in social networks (Mohseni and Williams 2019). Discrimination bias and conformity bias appear to be exacerbated by their expression through social networks. In the task of analysing what happens to human weaknesses through their digitalisation, social epistemology can contribute. In doing so, it can feed the ethics of information with remarkable insights that are relevant to its primary task of promoting an emancipatory social science. How can this goal be achieved? I will argue for an approach which has two aspects: i) the diagnosis and critique of the processes that produce these inequitable outcomes are an integral part of the philosophical framework of information technology; ii) this is conducive to making values explicit and including them as a feasible benchmark for the digital translation of actions. In addition, this process provides objectivity (Anderson 1995). That information technology embraces its value-ladenness is the backdrop to this framework. At the same time, this is both an emancipatory move and one that promotes objectivity. Why choose non-ideal theory over ideal theory? There are a number of good reasons to

move to non-ideal theory in information technology. The first reason is that an ideal theory would not cover the transformative effects of techno-social systems on work, organisations, industries, and society. A second, simple reason would be that in order to understand what fairness, equality and justice are, it is better not to abstract from the actual world in which we live. A reflection on idealised cases cannot possibly provide us with the proper basis for thinking about how to support fairness, justice, and equality (Doris 2015).

6. Bias and conformism make fact/value entanglement even more complicated

The empirical orientation of research on moral values in information technology does not mean that the prescriptive force of values (normativity) can be explained by science alone in naturalistic terms (Davidson 1970). Appeals to values are 'thick'. In ethics we speak of 'thin' and 'thick' concepts (Williams 1985). The use of thick concepts involves both non-evaluative description, i.e. various facts about their subject, and evaluation of their subject by reference to certain standards. In addition, values ask us to evaluate the role that their substantive claims play in the way we deal with the world and with ourselves (Nussbaum 1995). This process of evaluation engages the resources of reflection as well as emotional responses. What is more, values are not confined to the realm of beliefs; they are a call to action. Smith states: 'having such opinions is a matter of finding ourselves with corresponding motivation to act'. This feature of moral judgement is referred to as 'practicality' (Smith 1994). If you are convinced of the goodness of your moral judgement formulated in terms of value standards, then your actions should be guided by that judgement. Here values have therefore a function of orientation and guidance for the action. It does not mean that values prescribe a particular course of action, although concern for the right thing implies that there are right answers. Finally, values are not only individual but also collective in nature. We can think of values as a continuum of strength, encompassing descriptive and prescriptive features, rational and emotional aspects, and providing us with motivation and direction for action. From a historical perspective, the entanglement of facts and moral values, already noted by J. Ortega y Gasset (1923), has recently been discussed by Anderson (2004), Kitcher (2011), and Axtell (2016). In summary, the entanglement argument claims that a position that would separate empirical adequacy and ethical sustainability with respect to values is indefensible. Implicit rationality is not

only epistemic. It does not aim at merely intellectual objectives. The epistemic endeavour of predictive algorithms is to improve action in terms of objective, quantifiable, physiological, and behavioural data. Within this knowledge process, an agential point of view is implanted (Longino 1980). This enhanced state of affairs will produce a different dynamic from that of science. While the dynamics of science is considered to be affected by a lack of human-centred constraints because it is entirely inward looking, the enhanced agential standpoint claims a freedom of inquiry that does not play a role in the technological agenda.

Looking at values from the perspective of techno-social systems, we can say that scientific and technological innovations dramatically affect the values that people hold and share. As a result, a new vantage point has been made available for consideration of the impact that information technology has had on the ongoing discourse on moral values. It is legitimate to say that it provides new evidence for the central assumption about facts and values. It is also legitimate to say that the unique set of values, violations of values, and remedies underlined by techno-social discourse also provides new ways to investigate whether access to the negative space highlighted by new technology can be conducive to better practices—both with regards to the discourse on values and the practices in which the latter are involved.Finally, it can clarify whether certain biases have distorted our moral thinking and altered what we owe each other (Anderson 2015). Given all the above, it is fair to state that information technology is crucial to gaining new knowledge about the thick concept of 'values'.

Then, in order to describe an action as either blameworthy or virtuous in an exhaustive manner, one must consider the values that inspired it. For example, practices aimed at finding new vulnerabilities in techno-social systems, testing the security of software and devices, are to a large extent almost identical to practices that we want to discourage when undertaken by 'the bad guys'. This shows how hard it has become to evaluate facts without keeping in mind their entanglement with values in the realm of technologically mediated practices (Wolff 2016).

7. Ways of remedying bias and conformism in techno-social systems

The philosophical framework of information ethics is dramatically shaped by the idea that it has to be of practical relevance by its very nature. For this

reason, there is a convergence between the required condition for the operationalisation of information technology and the debate that takes place in political theory about the approach and the question of whether it is inspired by an ideal or a non-ideal theory (Stemplowska and Smith 2012). From this perspective, it is remarkable that information technology offers a novel understanding of the techno-political dimension that does not fit into the traditional conceptual cartography of the 'ideal versus non-ideal theory' methodological debate (Valentini 2012). It provides an original contribution to the debate on the practicality of politics and, more generally, of ethics. It explains the origins of values, adapting them to the results of the analysis of bias, and proposes remedies for impaired values.

The philosophical framework refers to the synergistic combination of three major disciplines, each of which is interested in practical outcomes, but has been kept separate from the others in the literature: (a) information technology; (b) ethics; (c) political theory (Sunstein 2017). I began by characterising the scope of the discourse on values in terms of practical attitudes. Now this initial characterization needs to be specified. Usually, however, the practical attitude is used in a much narrower sense than that explored here. It amounts to making judgments about the moral quality of an action (Strawson 1962, Smith 1994, Korsgaard 1996). In the digital space, moral judgement is only one activity among a wide range of other activities. Such a practical attitude has a broad scope, for it must also be understood in terms of finding ways to correct biases in techno-social systems. Actions to correct these biases have the goal of training techno-social systems and restoring threatened values.

One thing is ethics by design in the engineering of techno-social systems, which aim to intentionally influence individual behaviour by building norms into technical devices (Leenes and Lucivero 2014) or algorithms (Kearns and Roth 2020). Another is to align technological tools with the ongoing cooperative social and experimental process of democracy, in which what we owe each other is a matter of negotiation (Bozdag and van den Hoven 2015). In this area, the path of embedded ethics has not yet been sufficiently addressed. Chatila et al. (2021) argue that in order to make the development and use of modern techno-social systems consistent with human rights and values, we need to overcome their epistemic opacity and tendency to produce biases and false answers. Such technologies are not only beneficial; they can also make us vulnerable, both as individuals and as democracies.

To be more precise, ethics by design can be further broken down into three dimensions:

1. Ethics *in* design: accountability, ability of the system to be explained, democracy;
2. Ethics *by* design, responsive to ethical concerns and reasoning, aligned with shared ethical values;
3. Ethics *for* designers, professional ethics, professional codes of conduct.

Ethics by design involves five steps: (i) specifying, (ii) reconstructing, (iii) probing, (iv) broadening, and (v) converging/aligning (Boenink 2013). These five steps are related to the three dimensions described above. The process that these steps are likely to launch is initially divergent - it tends to promote reflection and gather expertise from the social sciences, ethics and political philosophy (Nurock et al. 2021). Once moral and political insights have been clarified, the final step is conducive to more practical outcomes.

8. The prescriptive force of values

Values have moral authority. How can we make sense of this claim? The initial explanation suggests that values have the power to morally require or forbid actions. However, they are not immune to objections; objections from reasonable points of view are relevant. According to the most plausible approach, the practices of punishment, blame, moral condemnation and approval are the basis of the prescriptive power of values (Strawson 1962). However, this is only the beginning of the story. These practices, the narrative continues, will evolve into judgements of their moral characteristics as good or bad. Asking whether values are *made* or *discovered* is one way to inquire about values and their authority. Putnam along with a number of other scholars asks this question: What are values? Where do they come from? His suggestion was that the better way to justify normative claims about values is provided by Strawson (Putnam 2002). Strawson, in fact, claims that we make up ways of dealing with problematic situations and that we discover which ones are better and which ones are worse. According to Strawson, normativity is pervasive in our lives, and an account based on reactive attitudes such as resentment and

blame can make this phenomenon easy to understand. As Korsgaard (1996) puts it,

> 'It is the most striking fact about human life that we have values. We think of ways that things could be better, more perfect, and so of course different, than they are; and of ways that we ourselves could be better, more perfect, and so of course different, than we are'.

However, this account of normativity in terms of reactive attitudes and their associated claims of blame or praise for ourselves or others needs to be weighed against qualified normativity. Usually this is done by reference to qualified standpoints. Some of them are expressed by ethical theories. According to normativism, agency requires the agent to recognize salient facts about themself, such as the reasons for their behaviour (Velleman 2000), that cannot be challenged by qualified points of view. Notably, this insistence on the fallibility of the human moral project is coupled with realism about values. In other words, the values are embedded in the form of life of the human being. They need to be evaluated because not every point of view is qualified. For example, many tweets from the Tay bot had to be deleted because they were offensive. This should be enough to reassure us that realism is not a subtle form of tyranny. Realism and fallibilism are part of the same story, just like the facts that belong to the development of science.

9. Conclusion

I argued that accounting for moral values in information technology would amount to making some progress in devising a new framework for the philosophical debate about values. In particular, information technology will take the philosophical debate about moral values to the next level. Since there is a lack of robust accounts of moral values, a more concrete approach that starts from a non-ideal situation with the goal of correcting the wronged values while providing better options for action could also benefit the science of values in our digitised society. In the same way that medicine saved the legitimacy of ethics, ethics can be revived by a closer look at what happens in the process of digitising some actions. This is not to say that our glorious experience is always something we should be proud of. In fact, when raw human data is used to train socio-technical systems, they acquire human biases. We feed them a lot of abuses that happen every day and are even depicted in

drama and fiction. However, in those cases, we can still judge the actions as evil and blame the characters for their wrongdoings. We can suggest how they might do things differently or be a different and therefore better person, or how they might be redirected toward a better choice of behaviour. Techno-social systems can only generalise what they have learned about our behaviour and therefore present us with our amplified discriminations, injustices, and biases. A closer relationship with actual human practices - to which this paper is committed - is not in itself a guarantee of alignment with our values. However, it may help to distinguish between human practices through which values-alignment is gained, and human practices through which values-alignment is lost. The exploration is not directed to the affirmation of values, but rather to the recognition of the abuse of values. Philosophy has begun to talk again about discrimination, bias, prejudice, and cognitive vulnerability as a result of information technologies. Abuses of values are not to be understood negatively by means of a prior understanding of the violated value. The genesis of the value discourse in information technology shows the opposite. The way to understand information justice is the opposite. Starting from what is very far from the rational ideal and very close to the common human experience, the concrete experience of moral values for human beings includes both sides of the story: the rational ideal and its violation. In this sense, digital injustice is the new normal that we must critically examine. In addition, hermeneutic injustice occurs when categorizations and ontologies lead to an inappropriate level of credibility towards a speaker's word. Even in this case, analysis reveals the uniqueness of what happens in digital environments. It may be useful to explain the case of regimes where the digital environment is just another tool to further consolidate the power of authoritarian states. For liberal democracies, socio-technical systems have a different penetration. Our collective hermeneutic resources are influenced by falsely objective and dispassionate ideals that seem to be better fulfilled by technological systems than by human judgement. Yet we may end up with epistemic behaviour that is less robust and less just. Early democratisation of the digital revolution can avert these kinds of undesirable tyrannical outcomes for our democracies.

REFERENCES

Anderson, E. (1995). Knowledge, Human Interests, and Objectivity in Feminist Epistemology. Philosophical Topics, 23(2), 27-58. Retrieved April 13, 2023, from http://www.jstor.org/stable/43154207.

Anderson, E. (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce". In *Hypatia*, 19(1) (pp. 1–24). doi:10.1111/j.1527-2001.2004.tb01266.x.

Anderson, E. (2015). Moral Bias and Corrective Practices: A Pragmatist Perspective. In *Proceedings and Addresses of the American Philosophical Association,* 89 (pp. 21-47). Retrieved April 13, 2023, from http://www.jstor.org/stable/43661501.

Axtell, G. (2016). *Objectivity*, Cambridge. Polity Press.

Battaglia, F. (2021). Truth, Knowledge, and Democratic Authority in the Public Health Debate. *HUMANA.MENTE Journal of Philosophical Studies*, *14*(40), 1-22. Retrieved April 13, 2023, from https://www.humanamente.eu/index.php/HM/article/view/377

Blanken-Webb, J., Palmer, I., Campbell, R.H., Burbules, N.C., & Bashir, M. (2019). Cybersecurity Ethics. In J.T.F. Burgess and E.J.M. Knox (Eds.) *Foundations of information ethics*. Chicago. ALA Neal-Schuman.

Boenink, M. (2013). The Multiple Practice of Doing 'Ethics in the Laboratory': A Mid-level Perspective. In van der Burg & Swierstra T. (Eds.) *Ethics on the Laboratory Floor*, London. Palgrave Macmillan.

Boucher, P., Nascimento, S. & Tallacchini, M. (2018). Emerging ICT for Citizens' Veillance: Theoretical and Practical Insights. *Science and Engineering Ethics* 24, (pp. 821–830). https://doi.org/10.1007/s11948-018-0039-z.

Bozdag, E. & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology* (pp. 249-265). 10.1007/s10676-015-9380-y

Brosch, T. & D. Sander (Eds.), (2015). *Handbook of value: Perspectives from economics, neuroscience, philosophy, psychology and sociology*. Oxford University Press.

Cappelen, H. (2018). *Fixing the language. An essay on conceptual engineering*. Oxford University Press.

Carens, J. (1996). Realistic and Idealistic Approaches to the Ethics of Migration. *International Migration Review*, 30 (1) (pp. 156-70).

Chatila R. et al. (2021). Trustworthy AI. In Braunschweig B., Ghallab M. (Eds). *Reflections on Artificial Intelligence for Humanity*. Lecture Notes in Computer Science, vol 12600. Springer, Cham. https://doi.org/10.1007/978-3-030-69128-8_2.

Davidson, D. (1970). Mental events. In L. Foster and J. W. Swanson (Eds.), *Experience and Theory*, (pp. pp. 79-101). Amherst. University of Massachusetts Press.

Doris, J.M. (2015). *Talking to Our Selves. Reflection, Ignorance, and Agency.* Oford University Press.

Elkins, J. & Norris A. (Eds.) (2012). *Truth and Democracy*. University of Pennsylvania Press.

Estlund, D. (2008). *Democratic Authority*. Princeton University Press.

Flanagan, M., Howe D.C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. van den Hoven, J. Wecker (Eds.) *Information Technology and moral philosophy*, New York. Cambridge University Press.

Frischmann, B., Selinger, E. (2018). *Re-engineering Humanity*, Cambridge University Press.

Haslanger, S. (2019). On the Epistemology of (In)Justice: Oppositional Consciousness and Social Critique. Haslanger Draft: 15 Oct 2019. Retrieved, April 13, 2023, from https://philosophy.cornell.edu/sites/phil/files/Haslanger%20Epistemology%20of%20Injustice%20for%20Cornell%20Law%20%20Phil.pdf

Haslanger, S. (2000). Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs*, 34: 31-55. https://doi.org/10.1111/0029-4624.00201

Huff C. & Cooper J. (1987). Sex bias in educational software. The effect of designers' stereotypes on the software they design. *In Journal of applied social psychology* 17 (pp. 519-532).

Hughes, T. (2004). *Human-built world: How to think about technology and culture*. Chicago. University of Chicago Press.

Hume, D. (1978). *A Treatise on Human Nature*, Oxford University Press.

Leenes, R.E. & Lucivero, F., (2014). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design. In *Innovation and Technology* 6(2) doi: 10.5235/17579961.6.2.193

Jasanoff, S. (2005). *Design on nature: Science and Democracy in Europe and the United States*, Princeton University Press.

Kearns, M. & Roth, A. (2020*). The Ethical Algorithm. The Science of Socially Aware Algorithm Design*. Oxford University Press.

Kitcher, P., (2011). *Science in a Democratic Society*, Amherst, NY: Prometheus Books.

Korsgaard, C. (1996). *The sources of normativity*. Cambridge University Press.

Kusch M, (Ed.), (2020). *The Routledge Handbook of Philosophy of Relativism*. London, New York. Routlegde.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. Bijker & J. Law (Eds.), *Shaping technology/building society*. Cambridge, MA. MIT Press, pp. 225–258.

Longino, H. (1980). *Science as Social Knowledge*. Princeton, N.J.: Princeton University Press.

MacKenzie, D., & Wajcman, J. (Eds.) (1985). *The social shaping of technology: How the refrigerator got its hum*. Milton Keynes, England: Open University Press.

McGinn, R.E. (1991). *Science, technology and society*, Englewood Cliffs, NJ. Prentice-Hall.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*. https://doi.org/10.1177/2053951716679679

Mohseni, A., Williams, C.R. (2019). Truth and Conformity on Networks. In *Erkenntnis*. https://doi.org/10.1007/s10670-019-00167-6

Moll, J. Zahn, R., de Oliveira-Souza (2015). The neural underpinnings of moral values. In Brosch, T. & D. Sander (Eds.). *Handbook of value: Perspectives from economics, neuroscience, philosophy, psychology and sociology*. Oxford University Press.

Moor, J.H. (1985). What is computer ethics? *Metaphilosophy*, 16 (pp. 266-275). https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Nissenbaum, H. (1998). The cutting edge. SIGCAS Comput. Soc. 28, 1, (pp. 38–39). DOI:https://doi.org/10.1145/277351.277359

Nussbaum, M. C. (1995). Aritstotle on human nature and the foundations of ethics. In Altham, J. E. J. und Harrison, R., (Eds.), *World, Mind, and Ethics*, Cambridge University Press, Cambridge, (pp. 86–131).

Ortega y Gassett, J. (1923). *Obras Completas*, Madrid, Revista de Occidente, vol. 6.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. New York: Penguin Press.

Putnam, H. (2002). *The collapse of the Fact/Value dichotomy*. Cambridge, London. Harvard University Press.

Railton, P. (1986). Moral Realism. *The Philosophical Review*, *95*(2), 163–207.

Shafer-Landau, R. (2009). *Moral Realism. A Defence*. Oxford: Oxford University Press.

Sayre-McCord, G. (2007). Moral Realism. D. Copp (Ed.) *The Oxford Handbook of Ethical Theory*. Oxford University Press.

Schroeder, M. (2007). Weighting for a Plausible Humean Theory of Reasons. In *Noûs*, 41 (pp. 110-132) https://doi.org/10.1111/j.1468-0068.2007.00640.x

Schwartz, S.H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*, Academic Press, Volume 25, (pp. 1-65), https://doi.org/10.1016/S0065-2601(08)60281-6.

Sen, A. K. (2000). *Development as Freedom*, New York. Anchor Books.

Smith, M. (1994). *The Moral Problem*. Malden: Blackwell Publishing.

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, Volume 48: 1962 (pp. 1-25).

Stemplowska, Z., Smith, A. (2012). Ideal and Nonideal Theory. In D. Estlund (Ed.) *Oxford Handbook of Political Theory*, Oxford University Press.

Stevenson, C. (1937). The Emotive Meaning of Ethical Terms. *Mind*, 46: 14–31.

Sunstein, C.R. (2007). *Republic.com 2.0*. Princeton: Princeton University Press.

Susser, D. & Roessler, B. & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2). doi: 10.14763/2019.2.1410.

Toulmin, S. (1986). How Medicine Saved the Life of Ethics. In *New Directions in Ethics*, Routledge.

Valentini, L. (2012). Ideal vs. Non-Ideal Theory: A Conceptual Map. *Philosophy Compass* 7, no. 9 (pp. 654–664).

van de Poel, I. (2018). Design for value change. In *Ethics of Information Technology* (pp. 1–5).

Velleman, J.D. (2000). From Self Psychology to Moral Philosophy. *Noûs*, 34: 349-377. https://doi.org/10.1111/0029-4624.34.s14.18

Verbeek, P.-P. (2019). Values that Matter: Mediation theory and Design Values, Academy for design Innovation management. In: *Research Perspectives in the Area of Transformations* Conference, London, pp. 396–407.

Whitaker, R.M., Colombo, G.B., Rand, D.G. (2018). Indirect Reciprocity and the Evolution of Prejudicial Groups. *Scientific Reports 8, 13247*. https://doi.org/10.1038/s41598-018-31363-z.

Williams, B. (1985). *Ethics and the Limits of Philosophy* Cambridge, MA. Harvard University Press.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1), (pp. 121-136). Retrieved April 13, 2023, from http://www.jstor.org/stable/20024652

Wolff, J. (2016) The hacking law that can't Hack it. Retrieved April 13, 2023, from *https://slate.com/technology/2016/09/the-computer-fraud-and-abuse-act-turns-30-years-old.html*