Ethical guidelines for the use of artificial intelligence and value conflict challenges

Thomas Søbirk Petersen

Roskilde University, Department of Communication and Arts, thomassp@ruc.dk

DOI: http://dx.doi.org/10.5324/eip.v15i1.3756

The aim of this article is to articulate and critically discuss different answers to the following question: How should decision-makers deal with conflicts that arise when the values usually entailed in ethical guidelines – such as accuracy, privacy, non-discrimination and transparency – for the use of Artificial Intelligence (e.g. algorithm-based sentencing) clash with one another? To begin with, I focus on clarifying some of the general advantages of using such guidelines in an ethical analysis of the use of AI. Some disadvantages will also be presented and critically discussed. Second, I will show that we need to distinguish between three kinds of conflict that can exist for ethical guidelines used in the moral assessment of AI. This section will be followed by a critical discussion of different answers to the question of how to handle what we shall call internal and external values conflicts. Finally, I will wrap up with a critical discussion of three different strategies to resolve what is called a 'genuine value conflict'. These strategies are: the 'accepting the existence of irresolvable conflict' view, the ranking view, and value monism. This article defends the 'accepting the existence of irresolvable conflict' view. It also argues that even though the ranking view and value monism, from a merely theoretical (or philosophical) point of view, are better equipped to solve genuine value conflicts among values in ethical guidelines for artificial intelligence, this is not the case in real-life decision-making.

Keywords: AI; ethical guidelines; algorithm-based sentencing; value conflicts

Introduction

In ethical discussions about the development and use of AI/machine learning, the industry (e.g. Google 2018), governments (e.g. European Commission 2020) and scholars (e.g. Bostrom and Yudkowsky 2014) argue that the development and use of these technologies ought to comply with certain ethical guidelines.[1] Just to give an example of how AI is used, throughout this article I will refer to the use of algorithm-based sentencing, which is common in many American States (Cohen

2015, Demulder and Gubbi 1983 and Freeman 2016). Algorithm-based sentencing (ABS) is an AI tool that is used to support a judge's decision when the judge needs to predict the risk of recidivism for an offender in the sentencing process. The algorithm behind ABS is fed information such as the offender's age, criminal record, sex, job status, religion, income, housing, family status and ethnicity. The algorithm is also fed a certain value for each of these bits of information in relation to the prediction of recidivism.

However, the authors of ethical guidelines for the use of an AI tool such as ABS do not always clarify what it means to comply with these ethical guidelines. In what follows, I adhere to the interpretation that in order to comply with such guidelines, the development and use of a certain application of AI – for example ABS[2] – should ideally not violate certain moral principles or values[3] entailed in these guidelines. We are here talking about values such as accountability, accuracy, prevention of harm, fairness, non-discrimination, respect for privacy, and transparency, which are all values that have a place in most ethical guidelines for the use of AI (see also Morley et al. 2019 for an overview of ethical guidelines for AI/machine learning technology).

Not only has important and detailed scholarly work been done to clarify the meaning of several of these values in connection with both the general use of algorithm-supported decision-making and, more specifically, the use of ABS. Ethical analysis of whether a specific use of ABS violates a specific value usually also follows from the conceptual clarification of each value. Herewith are some examples of elaborations concerning whether or when the use of ABS, or algorithm-supported decision-making in general, is, from a moral point of view, accountable (see e.g. Fink 2018 and Binns 2018a), sufficiently accurate (see e.g. Washington 2018 and Ryberg 2020), fair (see e.g. Friedler et al. 2016, Feller et al. 2016 and Binns 2018b), and transparent (see e.g. Carlson 2017).

Besides this detailed and highly clarifying analysis of each value, which is often undertaken with no detailed considerations or comparison of other values, little work has been done to clarify when a specific use of AI such as ABS has succeeded in following relevant ethical guidelines that typically involve several different values (such as those mentioned above). However, the latter type of clarification may seem redundant, as the obvious answer to the question of when ABS has been used in an ethically acceptable way is when its use has not violated any of the relevant values.

However, as some scholars have pointed out, values such as the above-mentioned may conflict with one another, in the sense that there are situations where we cannot comply with one value without violating another. Imagine, for example, that the supportive use of ABS by a judge, who needs to predict the risk of recidivism for an offender in the sentencing process, turns out to be more accurate than predictions made by judges who do not use ABS as a supporting tool for predictions of recidivism. Imagine also that, at the same time, it is clear that it is not transparent precisely which features (e.g. race, religion, income, housing and ethnicity) are fed into the algorithm and how these features in the algorithm are weighted in order to produce an outcome about the risk of recidivism of an offender. Imagine, furthermore, that the company or organization that has produced the algorithm wants to keep the features fed into the algorithm a trade secret in order to safeguard its profitability. Therefore, the company will not spend time developing and selling their more accurate algorithmic tool if transparency

concerning all these features is required. Finally, no better alternative exists, meaning that access to a more transparent and equally accurate algorithm to predict future recidivism of offenders is not available. In such a case – let us call it *the accuracy vs transparency case* – accuracy and transparency are in conflict, in the sense that you cannot comply with one without violating the other.

The aim of this article is to articulate and critically discuss different answers to the following question, raised by cases like those mentioned above. How shall we deal with these kinds of conflicts if values entailed in ethical guidelines for the use of AI often do or can conflict with one another or with other important values (that may not be part of the ethical guidelines in question), and yet we also want ethical guidelines to guide us when morally evaluating the use of AI tools?

This article is relevant for several reasons. First, there is no doubt that trying to understand and discuss how we should handle conflicts between different values – each of which should ideally not be violated – that are involved in the use of algorithm-based decision-making, like ABS, is of clear professional and academic interest.[4] This is especially so because few scholars working within the ethics of AI seem to recognize the problem of conflict between these values[5] or to have addressed it in any detail. Second, ethical analysis of, for example, the implementation or increased use of ABS to improve decision-making in the sentencing process should also be of interest to politicians, judges, victims, offenders, relatives of victims and offenders, and the general public, because as politicians and judges, they have to decide whether the criminal justice system should use ABS, or whether it would be morally acceptable but not obligatory to use ABS. Third, the consequences of using AI, such as ABS, will affect not only the sentencing process but also the outcome of the sentencing process. Both the process of sentencing and its outcome will usually have a huge impact on the well-being of those whose lives are most directly affected by the use of ABS – notably, offenders, victims, and the relatives of both. Finally, while it is necessary to clarify each value – such as accuracy, non-discrimination, transparency and privacy – and to investigate when the application of AI complies with each of these values, these types of investigations are not sufficient to determine whether the use of a certain kind of AI, such as ABS, is morally acceptable. Moreover, as already hinted, these kinds of investigations are not sufficient to decide, from a moral point of view, when we should accept a specific use of AI, as conflicts between values within ethical guidelines can often arise.

How should decision-makers (e.g. politicians and judges who are going to decide whether or how to implement ABS in criminal courts) deal with value conflicts that may arise from the ethical guidelines concerning the morally acceptable use of AI? In order to prepare the ground for answering this question, I will proceed as follows. In the first section, I want to justify why it is worth spending the time to discuss ethical guidelines. More precisely, I shall try to clarify some of the general advantages of implementing such guidelines in an ethical analysis of the use of ABS. Some disadvantages will also be presented and critically discussed. The purpose of the section to follow is to show that we need to distinguish between three kinds of conflict that can exist for ethical guidelines used in the moral assessment of AI such as ABS. Finally, I will distinguish between and critically discuss three different strategies when faced with what I call *genuine value conflicts*. The strategies are: the 'accepting the existence of irresolvable conflict' view, the ranking

view, and value monism. This article defends the 'accepting the existence of irresolvable conflict' view. It also argues that, even though the ranking view and value monism, from a merely theoretical (or philosophical) point of view, are better equipped to resolve genuine value conflicts among values in ethical guidelines for AI, this is not the case in real-life decision-making.

On some advantages and disadvantages of using ethical guidelines

Ethical guidelines for the use of AI, such as ABS, which incorporate moral values such as accountability, accuracy, non-discrimination and transparency offer several advantages. In the following discussion, I will focus on two of them: a) identifying *what values to comply with* when it comes to the use of ABS (e.g. via a checklist) and b) providing material for *critical moral reasoning* concerning the use of ABS.

One of the first steps in a thoughtful moral evaluation of the use of AI, such as ABS, is to be clear about which values political decision-makers, AI-ethicists, companies and the general public believe are important to take into account in order to estimate the moral status of a certain use of AI. Values expressed in ethical guidelines concerning the use of ABS can thus direct our attention towards the relevant values. Indeed, few scholars or politicians would say that we should not care about values like accountability, accuracy, non-discrimination and transparency. Instead, some researchers have stated: 'This list of criteria [in ethical guidelines for the use of AI/machine learning] … serves as a small sample of what an increasingly computerized society should be thinking about' (Bostrom and Yudkowsky 2014). However, the mere existence and knowledge of these values in itself does little to engender critical reasoning about the use of AI. This brings us to the next point.

Second, by directing our attention to such guidelines, we can equip ourselves for the critical reasoning needed in the moral evaluation of a certain use of AI, such as ABS. Consulting the guidelines and trying to estimate to what extent a specific use of AI complies with them can make us stop and think and discuss whether this particular use is morally acceptable, morally unacceptable or morally obligatory. As Morley et al. (2019: 3) have argued, for example, such guidelines can be used as normative constraints for the dos and don'ts of algorithmic use in society. As an example of this, we can refer to the software company Northpoint (now Equivant), which has created an algorithm called COMPAS. COMPAS is used nationwide in the USA to decide 'whether defendants awaiting trial are too dangerous to be released on bail' (Feller et al. 2016: 1). Northpoint refused to disclose the details of its proprietary algorithm, although it is claimed that the algorithm used is biased against African Americans. It is understandable that Northpoint refused full disclosure of the algorithm as it is a trade secret, and the company needs to protect its bottom line. However, this case at least raises ethical questions about relying on for-profit companies to develop risk assessment tools, if the use of ABS violates the value of transparency. Therefore, the move from identifying important values to critical moral discussion is important if we want to harvest the possible benefits of ABS and minimize the possibility of initiating 'ethical scandals' in the wake of ABS use.

However, before we discuss so-called conflict problems in the following sections, we can challenge the existence and use of ethical guidelines in some

obvious ways. First, there is the challenge that the mere existence of such guidelines would lead to *ethics shopping* or alternatively to *ethical whitewashing*. Roughly speaking, ethics shopping can be specified as the practice of choosing certain ethical principles from a variety of such principles in order to justify one's behaviour a posteriori, 'instead of implementing or improving new behaviours by benchmarking them against public, ethical standards [e.g. ethical guidelines concerning AI]' (Floridi 2019: 186). Ethical whitewashing (or *ethical window dressing*) can be defined as the practice of making 'misleading claims … or implementing superficial measures' in order to defend one's practice and thereby make it appear more ethical than it is in reality (Floridi 2019: 186).

Although these are serious challenges, I will not deal with these and related issues in any detail since several scholars within numerous fields, such as computer ethics and business ethics, have already tried to answer how we should handle the types of challenges which ethical guidelines are subject to (see e.g. Bowie 2017 and Floridi 2019). The solution concerning ethics shopping, according to Floridi, is to establish a shared, clear, authoritative and publicly accepted standard of what counts as morally good AI (Floridi 2019: 187). A big improvement is this direction is the European Commission's *Ethics Guidelines for Trustworthy AI*, with which anyone in the EU ideally ought to comply, according to the guidelines (European Commission 2020).[6] The solution to ethical whitewashing, according to Floridi (2019) and others, is transparency and education. Transparency clarifies whether the development and use of, for example, ABS do not violate values such as accuracy, accountability, non-discrimination and respect for privacy. A further strategy for minimizing ethical whitewashing is educating the public and politicians and having established professional ethical boards to deal with values and whether the values in the guidelines are actually implemented in the development and use of AI tools such as ABS. However, even if ethical guidelines do not lead to ethics shopping and do not function merely as window-dressing, problems remain for the implementation of ethical guidelines.

Apart from the above-mentioned criticism, the traditional philosophical criticism of ethical guidelines is based on the premise that the values mentioned in such guidelines are described in very vague terms, and furthermore, can be in conflict with one another. Moreover, insofar as this is true, such guidelines may therefore not provide much support for us when it comes to guiding our actions. If it turns out that guidelines cannot guide us, they are not of much use in the ethical evaluation of AI such as ABS. In the next section, I will briefly elaborate on the problem of vagueness of concepts and conflict between values. This work is important in order to set the stage for the final section, in which a critical discussion of several answers and solutions to the challenges of value conflicts will be the subject of inquiry.

The challenges of conflicting values

Next, I suggest that we differentiate between three kinds of possible conflicts between values when we want to apply ethical guidelines in the ethical evaluation of the use of an AI tool such as ABS. These three kinds of conflicts I call *conceptual conflicts*, *internal conflicts* and *external conflicts*. Let me present them and provide

examples, and then make clear why we will primarily discuss different kinds of answers for resolving the latter two kinds of conflicts.

A conceptual conflict arises within one single value and involves different interpretations of the single value in question. As an example, take the value 'non-discrimination'. If we want to know whether the data *provided by* an ABS are discriminatory or whether the *use of these data* is discriminatory, we need to know what we mean by the term 'discrimination'.[7] However, as Kasper Lippert-Rasmussen (2013) and others (e.g. Binns 2018b) have made clear, we can distinguish between different forms of discrimination. An important distinction in specifying the term 'discrimination' is the distinction between *direct* and *indirect discrimination*. One way to clarify what is meant by direct discrimination is to underline that it involves some kind of negative mental state or attitude – for example hate, hostility, stereotyping or neglect – towards the object of discrimination – let's say, African Americans. However, even though the development and use of ABS do not involve direct discrimination of African Americans, they may involve indirect discrimination of them. This observation has been a cornerstone of the critique of the Northpoint company that created the algorithm of the recidivism predicting tool COMPAS. For even if we assume COMPAS was produced with no negative mental states about African Americans (which I cannot confirm) and did not enter race as a feature in the algorithm (which is true), the output of the algorithm has been claimed to be biased against African Americans through indirect discrimination (Feller et al. 2016). This is because some of the other features fed into the COMPAS algorithm, such as low income, housing in high crime areas, unemployment and criminal records, result in more African Americans than European Americans being predicted to reoffend.

In the philosophical literature, we can find several examples of how different interpretations of the same value entailed in ethical guidelines can lead to conflicts in estimating whether an algorithm is transparent or not (Ryberg 2020 and Ryberg and Petersen 2021) or fair or not (Binns 2018b). The existence of these types of conflicts makes it difficult to know whether an algorithm is sufficiently transparent or non-discriminatory, for instance, or to compare whether one proposal for non-discriminatory algorithms is better than other proposals.

One obvious solution to this conceptual conflict is to try to specify what understanding of 'discrimination', 'transparency' or other value one wishes not to violate in the development and use of ABS. Although specifying our understanding might appear to be an obvious solution, this is not necessarily so. This is both because each value can be interpreted differently and because, from an epistemic point of view, it may be difficult to know whether a producer or a user of ABS is discriminating directly against a particular group of people, since it can be difficult to access the mental states of a person. However, by specifying a value like non-discrimination, we at least know where to look in order to determine if an algorithm or its use complies with a value such as non-discrimination or not.

The two other kinds of conflict – internal and external – are less discussed in the literature concerning ethical evaluations of AI and uses of it, such as ABS. Unique to these two kinds of conflict is that they arise not within one value, but between different kinds of values. For example, imagine a case where an algorithm supporting judges in predicting the recidivism of an offender in order to measure out the appropriate sentence is clearly constructed in a way where both the

construction of the algorithm and its application are both directly and indirectly discriminatory against women. In such a case, the value of transparency is satisfied, but the value of non-discrimination is not. Compare this with a case where it is not fully transparent how an algorithm for ABS has been constructed, but where the outcome of sentencing is less racist and sexist compared to decisions made by judges who are not supported by ABS. In such a case, at least the value of indirect non-discrimination may be satisfied but the value of transparency is not. We therefore need to know what to do in a conflict between values or at least clarify that a conflict exists, and decide whether it makes sense to say that a specific use of ABS complies successfully with the values embedded in the relevant ethical guidelines when we cannot have it all, so to speak. Parallel kinds of conflicts can occur if some of the values listed in the ethical guidelines clash with other values that are not listed there. A possible scenario might be if the value of privacy or transparency is satisfied in the development and use of ABS, but the outcome results in terrible consequences for the well-being of potential victims, and the value 'increasing well-being for the victim' is not part of the guidelines.[8] A certain use of ABS that would respect privacy, transparency and non-discrimination, for instance, might offer bail to many offenders, with the consequence that there could be many more victims of crimes than if more offenders were denied bail.

Therefore, if we accept that both internal and external conflicts of values do exist, it is of practical importance to know how we should deal with this if we want to apply ethical guidelines in the moral evaluation of AI decision tools such as ABS. One could argue here that if there is a conflict between certain values in the application of AI, then it should not be used. But the problem with this kind of reasoning is that if we do not allow conflict between the values in ethical guidelines for AI, then we should probably never use AI at all. Few would accept this decision since AI can be of tremendous help in solving or making us better equipped to handle big data and important social challenges in modern society.[9]

Strategies for dealing with value conflicts

So what are we to do when the values used to guide the use of AI come into conflict? This question can be interpreted as a more concrete version of a traditional and general question within normative ethics that arises for all ethical theories that accommodate the existence of more than one intrinsic value. In addition, I shall therefore draw from some of this literature on the subject to answer our more concrete question. I will present and critically discuss different kinds of strategies that we can call *the 'accepting the existence of irresolvable conflicts' view, the ranking view* and *value monism*.

However, before we move on to the investigation of these three strategies, we need to keep in mind that when we talk about conflicts between values, we are referring to what we could call *genuine conflicts* rather than *superficial conflicts* between values. Genuine conflicts between values are those that, unlike superficial conflicts, cannot be resolved by a deeper understanding or specification of the values in question. One example of a genuine conflict between values might be the conflict between accuracy and transparency mentioned in the introduction. A conflict of values leads us to perform incompatible actions: each value obligates us to do something we can do, but we cannot fulfil both values; that is, if we fulfil one,

we cannot fulfil the other.[10] On the other hand, a superficial conflict is an alleged conflict between values that evaporates when we specify the contents of the values in question. An example of how a conflict between values could be a superficial conflict is the case mentioned in the introduction involving accuracy versus transparency. In this case, we could argue that no conflict between accuracy and transparency exists, because we believe that satisfying the value of transparency does not mean that all the features fed into the algorithm by the (Northpoint) company ought to be transparent to everyone.[11] When superficial conflicts between relevant values are thus resolved we can say that this kind of use successfully complies with the ethical guidelines in question.

Therefore, if the conflict does not evaporate in attempting to specify or interpret the value in question, we have a genuine conflict. However, given this distinction between superficial and genuine conflicts, how can and should we handle a genuine conflict between values? I present and discuss three different strategies to address this question.

The first strategy is simply to explicitly accept that such conflicts may be irresolvable and that successful compliance with ethical guidelines for a specific use of AI may involve such value conflicts. Keeping the above-mentioned accuracy versus transparency case in mind – *it is not morally better (or worse)* to choose an algorithm that is more accurate but less transparent than it is to choose an algorithm that is more transparent but less accurate, given the available alternatives. This kind of reasoning is reflected in the work of bioethicists Tom L. Beauchamp and James F. Childress, who formulate their view as follows: '[value] conflicts sometimes produce irresolvable moral dilemmas. When forced to a choice, we may "resolve" the situation by choosing one option over another, but we may still believe that neither option is morally preferable' (Beauchamp and Childress 2013: 12). Another example of a philosopher who accepts that moral conflicts can be irresolvable is Rosalind Hursthouse, who writes: 'it is true both that a virtuous agent would do A, and that a virtuous agent would do B … both A and B are, in the circumstances, right' (Hursthouse 2013: 650). In sum, irresolvable moral conflicts can arise in situations where none of the possible decisions are morally preferable (this is the position of Beauchamp and Childress) or when several different decisions may all be morally right (this is the view of Hursthouse). Apart from these differences, we should accept that both adhere to versions of the 'accepting the existence of irresolvable conflicts' view.[12]

However, some obvious challenges to this view come to light.[13] The first challenge is epistemic. How do we know which conflicts are irresolvable? What if one group of people says that, from a moral point of view, we should accept a certain use of AI because transparency carries greater weight than accuracy, and another group of people reaches the opposite moral conclusion? Is such a conflict between these two groups of people irresolvable or not? Defenders of the view that certain conflicts between values are irresolvable do not say much about how we can know or differentiate between irresolvable conflicts and moral conflicts that are resolvable. What Hursthouse does say, however, is that if we accept the existence of irresolvable moral conflict, this should not be taken as a counsel of despair or used as an excuse for moral laziness. Adding to this, Hursthouse writes that '[i]t will always be necessary to think very hard before accepting the idea that a particular moral decision does not have one right issue' (Hursthouse 2013: 650). Again, this

does not move us much further along in discovering whether we are confronted with an irresolvable conflict or not. What does it mean to think hard about a moral question? Although each of us has some idea of what it means to think hard about a moral question, we may still differ considerably over what that means. For instance, when we consider what kinds of facts are relevant, for example in the moral evaluation of ABS, such considerations could easily be a new point of disagreement about whether we have taken into account a suitable number of empirical considerations or research studies or whether we have explained central concepts in a satisfying way. In addition, when have we thought hard enough about a moral question? After one hour? Or after having studied 10 ethical papers or reports about different answers to the given question? It is difficult to say anything precise about this, but that may not be necessary. When it comes to the moral evaluation of whether or not we should use an AI, if the society we live in has ideally gone through a democratic deliberative process in which all the parties affected by the technology have been heard and all the values mentioned in the introduction have been discussed among scholars, in public as well as in parliament, this rough description should satisfy the notion of 'thinking hard'.

A second challenge, this one of a practical nature, is the following: if irresolvable conflicts between values have been detected, then what? We cannot just settle on the view that we agree there is a conflict when it comes to the ethical evaluation of AI, as some policy is needed. What course of action is most appropriate? Should we just toss a coin?[14] Or should we consult a democratic process where the ethical view of scholars, politicians and those directly affected by the use of AI should guide our AI use? Will Kymlicka seems to accept the latter, when he reasons about the conflict between different moral values/principles in the context of assisted reproduction:

> Some potential conflicts between principles cannot be eliminated … In these circumstances, the Commissioners will have to balance the competing values as best they can, giving due weight to each. This is similar to the process of balancing values that judges are often confronted with. In both contexts we have a rough sense of when the process is being carried out impartially, and when someone is unduly biased towards particular interests. (Kymlicka 1993: 15)

While this kind of answer starts to clarify what we should do, it raises at least one challenge given that there are irresolvable conflicts. This challenge is practical in nature, and involves the very process of handling these conflicts. What does Kymlicka mean by 'balancing' or 'due weight', for example? Although Kymlicka does make some effort to clarify these concepts, writing that we should 'ensure … that all recommendations are checked against a comprehensive list of stakeholders and principles' (Kymlicka 1993: 26). However, although this statement does make us a little wiser, some challenges remain. What does it mean to let decision-makers considering the use of, for example, AI 'check' a recommendation against values like transparency and accuracy? Is it enough to show that you have considered all relevant stakeholders and values, by arguing why and how a specific use of AI does not violate a value like transparency but does violate the value of accuracy, and that based on these considerations, you have decided to accept the specific use of AI in question? This could at least be one answer. What Kymlicka suggests is better than ethical reasoning that is not guided by ethical principles or guidelines, and it aligns strongly with some of the advantages of having ethical guidelines, as mentioned in

the above-mentioned section. Dissatisfied with the 'accepting the existence of irresolvable conflicts' view, a common strategy for moral philosophers has been to defend strategies by which we can – at least in theory – resolve all or some value conflicts.[15]

A first strategy in favour of the view that moral conflicts are resolvable is to insist on a hierarchy among the values to be applied in a given context. This is the *ranking view*. One possible way by which a hierarchy could resolve moral conflict between values would be to rank these values according to their moral importance. Such a ranking could be done in numerous ways: for example, as done by John Rawls (1999), who stated that the satisfaction of one set of basic rights or values (such as freedom of religion, freedom of speech, the right to property or the right to vote) may never be violated for the sake of some other values (such as those expressed by his principle of fair equality of opportunities or the difference principle). However, in the literature on the ethics of AI, I have not come across any detailed hierarchy for values such as accountability, accuracy, non-discrimination and transparency described by any scholar or organization. The only hint of such a hierarchy I have been able to find is in the 2020 document by the European Commission which states that:

> AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. (European Commission 2020: 1)

The European Commission makes clear that in order to achieve this end, we should ensure that the ethically acceptable use of AI is based on values such as accountability, accuracy, democracy, respect for privacy, transparency and safety. However, the quotation above suggests that a hierarchy of values consists of three important values (increase human flourishing, bring progress and innovation) that should trump all others. Accordingly, the statement suggests that it would be morally right to violate a value (or values) like transparency and privacy if it turns out that the violation of the value (or values) in question would increase human flourishing. This value of increasing human flourishing may look like utilitarianism in disguise, but this is not the case. Utilitarianism tells us to maximize well-being, not just to increase well-being. The following case illustrates the difference. Imagine that you can assist only one out of three people by a certain use of AI. If you assist A, you will, all else being equal, not increase A's well-being. If you assist B, you will, all else being equal, increase B's well-being by one unit of well-being. If you assist C, you will, all else being equal, increase C's well-being by two units of well-being. In sum, you will increase well-being by assisting either B or C. However, you will maximize well-being only if you assist C. Another reason why this quotation by the European Commission on AI is not utilitarian is that it entails other values than increasing human flourishing, such as 'bringing progress and innovation'. However, although this is an example of a hierarchy of values that can resolve conflicts between some values (e.g. conflicts between the three values mentioned in the quotation and the other values such as accuracy, transparency and privacy mentioned in the guidelines), this hierarchy cannot resolve all value conflicts. It cannot resolve conflicts between the three values (increase human flourishing, bringing progress and innovation) or between values such as those listed in the guidelines (e.g. accuracy, transparency and privacy).

However, a third possible strategy to use when faced with conflicts between values is value monism in cases when the 'accepting the irresolvable conflict' view or the ranking view prove dissatisfactory. An example of a monism could be the value theory adopted by adherents of utilitarianism. If you are a utilitarian, moral conflicts can always be resolved, at least in theory. Simply choose the action – e.g. the use (or non-use) of AI – that maximizes well-being. Alternatively, if you do not know or are in doubt about what the consequences of AI use are, then choose the use of AI that we have good reason to expect will maximize well-being, even though it might violate one of the moral values such as transparency or privacy. If you apply this strategy, you will have succeeded in satisfying the ethical guidelines for AI.

Monism, at least in theory, can solve conflicts between values (because according to monism there is only one value), but it can also be challenged. First, one could argue that monistic theories such as utilitarianism themselves give rise to value conflicts, since there exists a degree of pluralism about what well-being is and how to measure well-being. The sheer number of different versions of theories of well-being, for example – ranging from hedonistic and desire-based theories to objective-list-theories – clearly indicates a conflict between different understandings of well-being. However, although this kind of conflict about how to specify the value of well-being only amounts to a conceptual conflict and not a moral conflict between the value of well-being and other values such as transparency, it still poses a potential conflict for decision-makers.[16]

Second, a further challenge to monism is that is does not (contrary to having a plurality of values) allow for the complexity and moral conflict of our moral experiences. Some philosophers (e.g. Beauchamp and Childress 2013) claim that it is an advantage that a pluralism of values reflects our moral reality. In addition, several empirical studies support the idea that people universally often adhere to several different values (see e.g. Haidt 2012) instead of just one. If this is true, and the universal public only takes seriously ethical guidelines that entail the values that the public adheres to, it is important that there not be too big a distance between public opinion and ethical guidelines. However, some challenges arise for this kind of criticism. First, although our current moral framework may be a good place to begin, history has often shown us that our morality has been wrong in several cases when we try to justify how we ought to act. Slavery and the sexual harassment of women are two prime examples. Second, adhering to a monistic value, such as maximizing well-being or human flourishing, does not mean that we should throw out all other values. Other values such as respect for autonomy, privacy, transparency and non-discrimination can serve as important instrumental values for the realization of the ultimate aim – which could be to maximize well-being. The quoted recommendation by the European Commission can easily be read like this. If this were the strategy, there would not necessarily be any distance between a utilitarian decision-maker and the public's adherence to several values in the evaluation of AI. This strategy works as a solution to the discrepancy between monism and people's opinions on values, but it does not resolve the challenge of what decision-makers should do when values are in conflict. The problem of values conflict has simply been shifted from intrinsic values to instrumental values.

A third challenge for monism is that monism, like utilitarianism, does not accept the existence of moral conflicts, as they can be solved in theory (e.g. by choosing the action that we can expect to maximize well-being). From a practical

standpoint, however, grounds for substantial moral conflict will probably still exist. This could be either because adherents of utilitarianism disagree about whether or not public transparency of the algorithm will maximize well-being, or because some are deontologists who believe that we must never violate transparency or respect for privacy, even if such violations would maximize well-being.

So, while value monism may look promising from a theoretical viewpoint, since it is simpler and more coherent than value pluralism, the acceptance of this view still involves many possible conflicts for decision-makers who have to deal with the ethics of AI in practice. Such conflicts are not only about issues that are conceptual (e.g. what do we mean by 'well-being' or 'autonomy') or empirical (e.g. will a particular use of AI, such as ABS, increase human well-being) but are also moral conflicts about what to do. However, it will not only be those moral conflicts that can arise if value monism accepts a plurality of values as instrumental, but what we morally believe to be right or wrong also depends on conceptual and empirical issues which may also cause conflict about what we ought to do.

When the ranking view and value monism do not seem to offer us much guidance in conflicts between values (since both approaches involve value conflicts and other inherent conflicts) in ethical guidelines, we are left with settling on 'accepting the irresolvable conflict' view. Although this view also faces some practical problems, it at least accepts that value conflicts need not be explained away and also offers some guidance about how to deal with real conflicts in a deliberative democracy.[17]

Conclusion

I hope to have shown that despite some advantages to using ethical guidelines in the moral evaluation of AI, such as ABS, one important and often overlooked challenge is the conflict that can arise between the values embedded in the relevant ethical guidelines. In dealing with this challenge we should first of all be clear about the kind of challenge we are referring to in order to be able to handle the conflict. Is it a conceptual challenge, or an internal conflict among the values expressed in the ethical guidelines, or an external conflict between values written into the guidelines and values that are not written into the guidelines? Is the conflict superficial or genuine? If we are confronted with a genuine conflict, we can use different strategies. First, we saw that the 'accepting the existence of irresolvable conflicts' view faces both epistemic and practical problems that could be reasonable resolved. Second, the ranking view, although a possibility for solving moral conflicts between values represented in guidelines for AI, has not been worked out by any scholars or governments in any detail. Even one of the best known examples of a ranking view developed by Rawls in his political philosophy clearly implies conflicts of values, too. Third, we discussed value monism as a strategy that, as least from a purely theoretical point of view, could resolve value conflict. However, analysing the workings of value monism revealed that the approach entails different kinds of conflicts that we could not expect to be resolved by those making decisions on the use of AI. From these discussions 'accepting the existence of irresolvable conflict' view appears to be the most realistic and the least problematic strategy for real life decision-making.

Notes

[1] According to Floridi (2019), more than 70 recommendations/guidelines about the ethics of AI have been published within the last couple of years (2017–2018).

[2] Instead of writing 'the development and use of AI' in what follows, for stylistic reasons I will mostly just write 'the use of AI'. This is not to neglect the point that ethical reasoning based on ethical values exists at each stage of the algorithmic pipeline. Moreover, this holds true for case development, design, building and testing through to use/deployment and monitoring. For a diagram of these different stages, see e.g. Morley et al. 2019.

[3] 'Moral values' is used here as a synonym for 'moral factors'. A moral factor determines the moral status of an act. So the overall moral status of an act, if it involves different values, depends on the interplay/weighting between different kinds of values or factors. For this kind of specification, see Kagan (1998: 17–18) – although Kagan only uses the term 'moral factors'. Moral values or moral factors concerning the moral evaluation of the use of AI can easily be translated into moral principles. Take privacy as an example: '… the AI system's impact on privacy and data protection, which are fundamental rights … which covers the respect for a person's mental and physical integrity' (European Commission 2020: 12).

[4] The problem of value conflicts is of course a general problem that confronts anyone who use ethical guidelines, whether it be within professions such as architecture, psychology, engineering or private companies.

[5] See e.g. Mittelstadt et al. 2016 who, when considering points of future research for algorithmic ethics, only focuses on the development of single concepts such as accountability, transparency, autonomy and privacy. The issue of how to handle conflicts between these values is not even mentioned.

[6] A further step towards harmonization and therefore against ethics shopping is also found in Section 2.4 of the *Ethics Guidelines for Trustworthy AI* 2019 (European Commission 2019), where it is mentioned that the EU wants to work in favour of building an international consensus on the ethics of AI.

[7] 'Discrimination' is conceived of here as a morally problematic factor, as this fits in well with the idea that non-discrimination is a value that should not be violated according to the ethical guidelines for the development and use of AI.

[8] The case mentioned here is constructed purely to illustrate a logical possibility. Usually, guidelines for AI do include the 'no-harm' principle or that 'the ultimate aim is to increase human well-being' (European Commission 2019: 1).

[9] I will only focus on examples of internal conflicts, as these are easily noticed in the scholarly and public debate on the use of AI. External conflicts can easily be translated into internal conflicts when those who have worked out the guidelines can accept that a value that is not presented in their guidelines should in fact be incorporated into them. At least, this is my own experience after working on several ethical councils for the state or for private companies.

[10] See e.g. Tännsjö (2011) for a specification of a value conflict close to my description.

[11] This way of reasoning about a possible solution to an alleged conflict between different moral values is inspired by Hursthouse (2013).

[12] A view similar to the 'accepting the existence of irresolvable conflicts' view is also presented and argued for by e.g. Williams (1981) and Stocker (1990 and 1997).

[13] See e.g. Mason (2018) for a splendid overview of arguments both for and against value monism as well as value pluralism.

[14] Hursthouse (2013: 652) writes explicitly that tossing a coin is not the morally appropriate way to solve a moral conflict – 'would the virtuous agent toss a coin? Of course not.' By why should a virtuous person not want to toss a coin? At least we can imagine situations where it would be the right thing to do – even for a virtuous person. If we are in need of a choice and if both choices are morally right – then it should not matter which one we choose. Moreover, there may be situations where it may be far more important that *we make a choice* than *how we make it*. Even though one finds it morally problematic to toss a coin in order to solve a moral problem, it may be morally more problematic to delay making a choice or to not make a choice at all. A virtuous agent should, for example, not delay making a choice between alternative X or Y if both alternatives, all else being equal, would promote well-being to the same extent, but for different groups of otherwise identical individuals.

[15] One obvious answer for how one can resolve moral conflicts, at least in theory, is to adhere to what is called *metaethical cognitivism*, according to which moral views and moral principles/theories can be true or false; see Parfit (2011) for the view that cognitivism can solve moral conflicts and Rowland (2021) for a critical discussion of this view. So if we have an alleged moral conflict between two people, with one person arguing for ABS use and the other against ABS use, it can always be resolved by appealing to the true morality of the matter in question. However, to discuss in any depth whether cognitivism is plausible or not would take far more space than we can allow in an article on how to deal with value conflicts within ethical guidelines for AI. Furthermore, even if cognitivism were true and could in theory solve moral conflicts based on value conflicts, that would not eliminate the existence of value conflicts in practice. Cognitivism would not be able to solve value conflicts between people having to make decisions. The reason for this is that not everyone would know about or accept cognitivism as true and/or because people would still disagree about which values are the right ones. Further reasons for why resolving value conflicts between people having to make decisions include differing opinions on how we should understand these values, for example, whether each value should carry the same moral weight, and what answers about moral issues we can or ought to derive from these values.

[16] A moral conflict among utilitarians about whether conduct C is right can of course also arise from differences in opinion about what the currency of 'utility' is, and what the expected consequences of the utility of C is estimated to be. The same types of problems can arise with monistic versions by deontologists who believe that a certain kind of conduct should not be violated (e.g. not harming or risking harm to innocent humans), because people have conflicting opinions about what counts as harm and what counts as a morally problematic risk of harm to others.

[17] For a further discussion of how to deal with such conflicts through compromise, see e.g. Kappel (2018) and Rowland (2021: Chap. 8).

References

Beauchamp, T. L. & Childress, J. F. (2013). *Principles of biomedical ethics* (7th edition). New York: Oxford University Press.

Binns, R. (2018a). Algorithmic accountability and public reason. *Philosophy and Technology*, 31: 543–556. https://doi.org/10.1007/s13347-017-0263-5

Binns, R. (2018b). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research* 81, 149–159, 2018 Conference on Fairness, Accountability and Transparency.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge, UK: Cambridge University Press.

Bowie, N. E. (2017). *Business ethics: A Kantian perspective*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781316343210

Carlson, A. M. (2017). The need for transparency in the age of predictive sentencing algorithms. *Iowa Law Review* 103, 303–329.

Cohen, M. (2015). When judges have reasons not to give reasons: A comparative law approach. *Washington and Lee Law Review 72*, 483–571.

Demulder, R. V. & Gubbi, H. M. (1983). Legal decision making by computer: An experiment with sentencing. *Computer/Law Journal* 4, 243–303.

European Commission (2020). *Assessment list for trustworthy artificial intelligence.* European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 21 December 2020.

Feller, A., Corbett-Davies, S., Pierson, E., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labelled biased against blacks. It's actually not that clear. *The Washington Post*, Oct. 17. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. Accessed 21 December 2020.

Fink, K. (2018). Opening the government's black boxes: Freedom of information and algorithmic accountability. *Information, Communication & Society* 21:10, 1453–1471. https://doi.org/10.1080/1369118X.2017.1330418

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology* 32:2, 185-193. https://doi.org/10.1007/s13347-019-00354-x

Freeman, K. (2016), Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law and Technology* 18, 75–106.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. arXiv:1609.07236v1 [cs.CY] https://arxiv.org/abs/1609.07236. Accessed 17 May 2021.

Google (2018). *Google's AI Principles* (by Sundar Pichai) https://www.blog.google/technology/ai/ai-principles/. Accessed 17 May 2021.

Hursthouse, R. (2013). Normative virtue ethics. In R. Shafer-Landau (Ed.), *Ethical theory: An anthology,* 2nd edition. Chichester, UK: Wiley Blackwell.

Kagan, S. (1998). *Normative ethics*. New York: Routledge.

Kymlicka, W. (1993). Moral philosophy and public policy: The case of NRTs. *Bioethics* 7:1, 1–26. https://doi.org/10.1111/j.1467-8519.1993.tb00268.x

Lippert-Rasmussen, K. (2014). *Born free and equal? A philosophical inquiry into the nature of discrimination*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199796113.001.0001

Mason, E., (2018). Value pluralism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 Edition). https://plato.stanford.edu/archives/spr2018/entries/value-pluralism/.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society* 3:2, https://doi.org/10.1177/2053951716679679.

Morley, J., Floridi, L, Kinsey, L, & Elhalal, A. (2019). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. https://doi.org/10.2139/ssrn.3830348

Ryberg, J. (2020). Risk-based sentencing and predictive accuracy. *Ethical Theory and Moral Practice* 23, 251–263. https://doi.org/10.1007/s10677-020-10066-3

Ryberg, J. & Petersen, T. S. (2021). Sentencing and the algorithmic transparency/accuracy conflict. In J. Ryberg & J. Robert (Eds.), *Principled sentencing and artificial intelligence*. Oxford, UK: Oxford University Press (forthcoming).

Tännsjö, T. (2011). Sophie's choice. In W. E. Jones & S. Vice (Eds.), *Ethics at the cinema*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195320398.003.0011

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal* 17, 131–160.