Ethical Algorithms

A brief comment on an extensive muddle

R.J. Anderson
Horizon Digital Research
University of Nottingham

W.W. Sharrock
Department of Sociology
University of Manchester

This paper was written while we were developing *Post Modernism, Technology and Social Science*. It didn't seem to fit that collection and so has a life of its own.

Introduction

Felicitas Kraemer and her colleagues want us to accept that some algorithms are ethical. They go to great lengths to explain and defend what they mean by this. Here is their initial definition:

> *Many, but not all, algorithms implicitly or explicitly comprise essential value judgments. By an 'essential value-judgment' we mean the following: If two algorithms are designed to perform the same task, such as classifying a cell as diseased or non-diseased, these algorithms are essentially value-laden if one cannot rationally choose between them without explicitly or implicitly taking ethical concerns into account. Another way of saying this is that the algorithm cannot be designed without implicitly or explicitly taking a stand on ethical issues, some of which may be highly controversial. (Kraemer 2011 p 251)*

A little later, they unpack their definition somewhat:

> *An algorithm comprises an essential value-judgment if and only if, everything else being equal, software designers who accept different value-judgments would have a rational reason to design the algorithm differently (or choose different algorithms for solving the same problem).(Kraemer 2011 p 253)*

We believe the use of definitions such as these will have the consequence of promoting ethical issues and challenges where none exist. In contrast to Kraemer et al, we propose a deflationary strategy. If you like, we suggest the adoption of a kind of Occam's Razor for ethics in ICT: "Do not multiply ethical issues beyond necessity". This would have the result, or so we think, of restricting focus to the very real and important challenges which do exist without our being unnecessarily distracted by figments of philosophers' imaginations. To justify our deflationary intent, we will offer four arguments which, taken together, show that the concept of ethical algorithms is a much overblown one.

Not all Value Judgements are Ethical

For the moment, let us set aside the issue of whether algorithms are the sort of thing to which we can attach the descriptor 'ethical' and assume for the sake of argument that we can. Kraemer et al are clear that if, because they held to different value structures, designers could have good reason to design the algorithm differently, it follows that the algorithm is an ethical one. The second definition we cite above states this very clearly. The first definition tells us that all such value judgements will result in ethical algorithms. In other words, the class of ethical value judgements is isomorphic with the class of all value judgements. Does this seem reasonable?

Value judgements, as their name suggests, are sets of evaluations, positive or negative (or even neutral). And, while it is true that philosophers have usually been concerned with ethical value judgements, it is by no means clear that Philosophy (or even sub-branches of Philosophy) hold that there are only ethical value judgements. In *The Place of Reason in Ethics*, for example, Stephen Toulmin suggests the term 'gerundives' for that groups of concepts which confer or convey value judgements.

> *The name 'gerundives' is appropriate because they can all be analysed as 'worthy of something-or-other; in this resembling the grammatical class of 'gerundives', which appears in one's Latin primer — consisting of such words as amandus, which means 'worthy of love' (or 'meet-to-be-loved'), and laudandus which means 'worthy of praise'. (1986, pp71-2)*

Toulmin suggests instances of such gerundives are logical and aesthetic value judgements as well as ethical ones. Following Hilary Putnam's lead, we would add 'epistemic' value judgements to the list. Different value judgements apply to very different forms of appraisal, as we very well know. Praising the well formed nature of an argument is not the same as praising the beauty of a painting or the elegance of a building. And neither is it the same as the endorsement of the rightness of a moral action. Epistemic value judgements pertain to claims about knowledge and in particular what will count as knowledge in science. Epistemic value judgements will be very important in one of our later arguments.

There are, then, many different kinds of value judgements each with their own provenance and (to use a somewhat outmoded phrase) their own logical grammar. Running the types together produces a kind of conceptual mash-up within which important distinctions are confused or even lost. Even if designators such as "good" are used in common, in regard to each domain analysis of what we mean by a 'good argument', 'a good book', a 'good action' will mean very, very different things. As Judith Jarvis Thomson remarked paraphrasing John Austin, good is like real in that it is substance hungry. To understand what it means in regard to any case, we need to know what the case is good for. There is not, as G.E. Moore thought, a common property characterisable as 'goodness'. How goodness manifests itself will vary according to the domain in which we are using the term.

There are two points to be made about all this. First, we can readily accept that (some or even all) algorithms might be constructed using value judgements. That does not *ipso facto* make

them ethical in nature.[1] Recognising the range of gerundives and the differences between them has the immediate effect of restricting the application of the description 'ethical' to algorithms where ethical value judgements have explicitly featured in their construction. This will significantly curtail the range of cases subject which be subject to review. Second, if Kraemer et al really do want *to engage* software designers and others in a reasoned debate on the subject of the ethics of algorithms (as opposed to simply haranguing them), having a clear and simple definition is going to be vital. It will not bode well either for the likelihood that the debate will be either reasoned or convergent if its central terms are loosely formed.

"Ethical Algorithm" as a Transferred Epithet

Although they use "ethical" as a global descriptor for what are no doubt both 'ethical' and 'unethical' algorithms (indeed, they are primarily interested in the latter, it seems), in using the term in relation to algorithms Kraemer et al are using the designation in much the same way as it is used in such phrases as "ethical products" or "ethical business practices" or "ethical investments" (substitute 'unethical as needed). "Ethical", then, modifies the term to which it is applied.

How does this modification work? How do "investments", "business practices", "algorithms" and so on get to be such that they can be called "ethical" or "unethical"? Kraemer et al provide a clear and detailed account of the process at work for algorithms, and no doubt similar or analogous processes are at work in the other cases. In brief, their account goes something like this:

1. A designer of any kind of artefact has to make a myriad of decisions in the process of completing the design. Design is about narrowing options and making choices.
2. Most of these design decisions will be derived from decisions about how to interpret the specification, or to conform to regulation, industry standards, or because of the character of prior decisions about types of material, ranges of function, and so on.
3. However, some of these decisions will be about features or properties of the artefact that might well reflect certain kinds of professional or personal pre-dispositions. (Kraemer et al call these properties 'contextual'). So, for example, the designer of a packaging product might specify that the product must be made from recycled materials, or be bio-degradable. Equally, the design of the labelling might specify the wording be printed in braille as well as English (or whatever).
4. For Kraemer et al, contextual decisions such as these are *ethical* and the artefacts so produced are ethical artefacts.

What are we being offered here? We have an account of design as a process of reasoned decision making. Some of the reasons for making such decisions might well be ethical (in the

[1] For example, Djikstra's *GoTo statements considered harmful* turns on an aesthetic and efficiency value not an ethical one.

Kraemer et al sense or in a narrower sense). But does this mean we can call product (in both senses) of those decisions "ethical"?

The use of "ethical" here is no more than a trope, a figure of speech, used for rhetorical effect. When travelling to the funeral of a loved one and reflecting on the life of the person who has died, we might well report that we had a "melancholy journey". Equally, feeling in a good mood in the morning, we might well say we had a "cheerful shower". The epithets "melancholy" and "cheerful" have been transferred from our reflections or sense of well being to the journey and shower. Journeys can be no more melancholy than showers can be cheerful. Of course, in ordinary discourse we know what we mean by these phrases. In using them, we are saying something about the way the journey was made or how the showerer feels or might feel. As descriptions of the journey and the shower they are perfectly acceptable metaphors. What they do not do is give us *additional or indeed any* factual information about the journey or the shower itself. The problems begin when we start to assume that such metaphorical description actually is a factual one (or a disguised factual one). It would be quite odd if someone were to respond to our description of our journey by offering to devise ways to raise the spirits of the journey. It would be even odder if they were to ask if the cheerfulness of the shower was infectious. Our listener would have failed to see what in using the trope that we did, we were saying about ourselves.

What this alerts us to is the role of this particular device in the persuasive strategy employed by Kraemer et al. If we allow that some algorithms simply are factually "ethical" in themselves, then the designation used in relation to them ceases itself to be an appraising one, and hence one over which we could hold differing (subjective? see below) opinions. That they are ethical is put beyond debate. What is then open for debate is just what their ethical character is. Are they blame- or praiseworthy? It is hardly a co-incidence that the leading example Kraemer et al use to illustrate the supposed ethical implications of the choices made is image processing in mammography. Choices about where to set thresholds certainly do have important consequences here in that tumours might be missed or unnecessary operations performed. The emotionally imbued character of the use context allows the ethical designation to be slipped in. Would they feel as concerned about number plate recognition?

The combination of the use of a transferred epithet and emotive examples is designed to secure acceptance of a rhetorical strategy not to demonstrate the fact of the matter.

"Ethical Algorithms" as Category Error

We have seen that "ethical" is used by Kraemer et al as a catch-all term for value judgements. We have further seen that when it is used, it is as part of an attempt to persuade us of the ethical (or unethical) use of such algorithms rather than as a purely empirical description. The term comes trailing clouds of rhetoric. The urge to persuade and the failure to distinguish result in a usage

which as we saw earlier perpetrates a category error in which entities, the algorithms in question, are attributed properties which they cannot have.

What do we mean by this? As Gilbert Ryle pointed out when he first introduced the term, category errors arise when people are...

> *...liable in their abstract thinking to allocate ... concepts to logical types to which they do not belong. (Ryle 1973 p 19)*

He illustrates the notion by such examples as the visitor to the university who is taken on a tour of the departmental buildings, the libraries, the student accommodation and so forth and who, at the end of the visit, asks to see the University as well; or of the small boy who watching solders marching by sees the units and battalions and asks where the army is. A University does not exist as an entity like a departmental building or a library. An army is not a group of soldiers over and above the units, battalions etc.

How does this apply in the Kraemer et al case? The logical grammar of the concept "ethical" refers to processes of reasoning and the grounds on which decisions are framed. Ethical decisions are based on ethical principles. The process of transferring the epithet from the decision to the constructed product enables us to think that there are *kinds* of algorithm; the ethical and the non-ethical which are just like the kinds of decisions we might take and the courses of action based upon them. We find the Good Samaritan's behaviour ethical and applaud it. We see how his intervention was borne out of compassion and a sense of duty towards the man who had fallen among thieves. "Ethical", then, is first a term of appraisal applied to social actors and their actions and not simply or only a descriptor (though sometime it is used as a descriptor) and second a term which is applied to courses of action taken in normatively defined contexts. The context determines if the action is to be considered *from an ethical point of view*. An action isn't ethical (or unethical) until we find it to be so. An algorithm is not a course of action in a normatively defined social context. At its simplest it is a formalised recipe; a set of instructions for repeatedly performing some task. As they are used in Mathematics, their rules of application, inputs and outputs are formally defined (that is, defined independent of the context in which they might be used). The inputs to an algorithm are variables; the outputs are functions. A social actor, say the designer or the user of the system, might specify a variable in order to have a specific normatively valued outcome. But that doesn't make the algorithm ethical. It is the action of the social actor which we might judge to be ethical or not. Just as universities are not entities in the same way buildings are, so algorithms are not actions in the same way that the Good Samaritan's were. To think that they are, is to commit a category error, in the same way as the person who wants to cheer up a melancholy journey up commits a category error.

The Enchantment of the Fact/Value Dichotomy

The reason that the notion of value can insinuate itself so easily into the discussion of the nature of algorithms is because of the continued fascination that many scientists and some philosophers have with the so-called "fact/value" dichotomy. According to the proponents of the dichotomy, a proposition (statement or description) cannot be both a statement of fact and a statement of value. The former describes "objective" properties of the matter in hand; the latter subjective opinions. Moreover, according to the prohibition often called Hume's Fork, it is not possible to argue logically from statements of fact to statements of value. The two discourses are distinct. Consequently, under this view, it seems reasonable to say that if one can show that some proposition is not factual (by whatever lights are accepted as determining that designation), it must therefore be held to express a value (or for Kraemer et al be "ethical").

The way this surfaces in relation to algorithms is through the following considerations. Breast cancer screening these days involves digital mammography. Given that large numbers of such images are taken every day and that 'eyeballing' each image to look for potential tumours is time consuming and relatively expensive, an initial triage of the images is performed by a scanning algorithm. In designing the scanning algorithms, designers of the software have sought to render in code ways of assessing the shape and texture of internal condition of the breast. Although these techniques have been improved and are being improved, they are not foolproof. Artefacts of the scanning procedure and process generate 'noise' in the image. Such noise introduces the risk of what might be thought of as 'Type I' and 'Type II' errors. The former are *false positives* whilst the latter are *false negatives*. Since it is impossible to eradicate all noise from the images, there will never be a completely error free scanning algorithm, though of course the probability of error can be and has been massively reduced. The question the designers must resolve is where to put the cut-off point. Should they predispose their system to generate more false negatives than false positives by being too strict in the definition of cancerous material, or vice versa? Or should they make both equally likely? Increase the likelihood of false negatives and you increase the likelihood of tumours being missed. Increase the likelihood of false positives and you increase the likelihood of women having unnecessary worry and/or potentially dangerous operations. The decision where to put the cut off, Kraemer et al tell us, is an ethical one and the algorithm, no matter where the balance is struck, is value laden.

Of course it is quite straightforward to see why Kraemer et al arrive at this conclusion. Undetected tumours and unnecessary operations are 'harms' of different kinds. As human beings and as medical practitioners, doctors are expected to avoid harming their patients or having their patients put at risk of harms. Where they do expose their patients to such risks, there have to be strong ethical grounds for doing so. Using the scanning algorithm as part of the triage process

exposes patients to risks of harms, therefore the algorithm and its in-built decision processes are value laden and hence ethical.

> *Software for complicated tasks such as medical diagnosis is often immensely complicated. Much if it consists of components that may have been developed earlier for 'general purposes'; these components are reused in order to make the software production process economically feasible. Segmentation algorithms are an example of such components. Components are preferably treated as black boxes: based on their formal, functional specifications, their behavior can be assumed to be 'correct'. The ethical position, however, that was adopted during the construction of such a component when either choosing a more conservative or more liberal threshold, is typically not part of their formal specification. That means that the same segmentation algorithm, applied in two different systems, will behave equally with respect to its formal, functional requirements, but at the same time it may behave oppositely with respect to its tendency to produce false positive or false negative judgments (Kraemer et al p 156-7)*

The paradigm case being used to assess algorithmically supported decisions is one that is based on the assumption that the facts are independent of (subjective) interpretation. If it is not possible to determine "the fact of the matter", then the decision can only be an (ethical) value judgement. Here (revealingly) is a comment attached to Figure 2 in Kraemer et al which shows the trace of a (noisy) image and where thresholds might be put.

> *For a larger threshold value, T1, the estimated area V1 of the 'blood' segment will be lower than the estimated area V2 which is found with a lower threshold T2. It is not a priori clear, however, which of the two threshold values is the 'correct' one. The software engineer chooses a threshold without real argument—thereby biasing the outcome of the algorithm when it is used on patient data. This may statistically influence the change of false positive diagnostic errors in favour of false negatives, or vice versa (Kraemer et al p152)*

Whilst it may not be possible practically to fix where the correct threshold should lie, in principle it should be. *And moreover there is just such a correct value*. John Austin once called words like "correct" 'trouser words' since in relation to their antonyms they 'wear the trousers' and determine meaning. Anything that departs from this sense of correctness (the just one way it should be) is "incorrect" and hence for Kraemer et al any judgment based on it a value judgement.

The metaphysics of the Fact/Value dichotomy (there are facts and there are values and never the twain shall meet) has come under intense scrutiny during the last 50 years or so. No-one has been more persistent in his criticism nor more persuasive in his arguments than Hilary Putman. Putnam has argued that investigations into logic and into the practices of science all lead to the conclusion that the fact/value distinction has collapsed. As we have moved away from representational theories of meaning posed by treating the individual as an asocial, Cartesian observer (how do my descriptions of states of affairs come to be true independent of me?) to a recognition that representation and description much as elsewhere are deeply embedded social practices in science, we have come to appreciate that the 'truthfulness' or 'factuality' of a proposition is itself based upon presumptions as to what will count as truth or fact in particular forms of discourse. These are the epistemic values we referred to above. Because there is no way to prise ourselves out of our (social and professional) milieux in order the measure the correctness of a proposition (there is no view from nowhere), how we determine truthfulness and factuality in technology, science and everyday life is by accordance with judgements of value in use in the relevant practices. It is the practices of science, technology, medicine etc as professional and collective endeavours that tells us what is true, factual, the case in science, medicine and technology. Such value judgements are judgements of what is expectable, reasonable, in accordance with experience and so on. We do not have facts on the one hand and values on the other; values underpin facts.

Saying this, though, does not mean we immediately slide down Kraemer et al's slippery slope. The objectivity of the judgements of truthfulness or factuality in science and so forth is guaranteed by the observable, reviewable, accountable use of the practices. Coming to these conclusions using practices such as this *is exactly what we mean* by objectivity. It is only because they unwittingly hold to an empiricist view of fact and value and its inherent representational theory of meaning that Kraemer et al can set off on their search to find the value base of (some) algorithms. The muddles and confusions they create on the way, allow them to arrive at their (implausible) destination.

Conclusion

The notion that algorithms and other computational artefacts might be ethical has gained ground recently. It has been particularly championed by Philip Brey, Lucas Introna and members of the Value Sensitive Design fraternity. The discussion which Kraemer et al provide of the idea is as clear a summary of what is meant as we have found. Its clarity is its saving grace, since because the discussion is so detailed it is possible to pin down just where the muddles are coming from. The outcome of allowing the muddles to continue can only be unholy and unruly contention rather than reasoned and reasonable debate. Were this to happen, it would be a disappointment and a shame. There are difficulties enough regarding the use of modern technologies in the world for us to work

on and many of them have deep and challenging ethical implications. We have no need to add well meaning muddles to the raft of those important and real ones we already have to hand.

References

Kraemer, F., van Overveld, K. and Peterson, M.	2011	Is there an ethics of algorithms? *Ethics of Information Technology*; 13:251–260
Toulmin, S. E.	1986	*The Place of Reason in Ethics*. Cambridge. Cambridge University Press.
Ryle, G.	1973	*The Concept of Mind*. Harmondsworth. Penguin Books.