Cognitive Mimetics for AI Ethics: Tacit Knowledge, Action Ontologies and Problem Restructuring

Antero Karvonen[1]

[1] University of Jyväskylä, Jyväskylä, Finland

antero.i.karvonen@jyu.fi

Abstract. Ethics and ethical information processing are an important problem for AI development. It is important for self-evident reasons, but also challenging in its' implications and should be welcomed by designers and developers as an interesting technical challenge. This article explores AI ethics as a design problem and lays out how cognitive mimetics could be used a method for its design. AI ethics is conceptualized as a problem of implementation on the one hand, and as a problem of ethical contents on the other. From the viewpoint of human information processing, ethics becomes a special case of *ethical* information processing - one that has deep implications in terms of AI abilities and information contents. Here we focus on ethical information processing as a property of the system (rather as a general constraint on it). We explore three specific concepts relevant for cognitive mimetics from the perspective of ethics: tacit knowledge, ontologies, and problem restructuring. We close with a general discussion on the difference between abilities and mental contents noted as relevant in previous articles on cognitive mimetics and reiterate its importance in this context as well.

Keywords: AI ethics, AI design, cognitive mimetics, mimetic design

1 Introduction

At least in theory, artificial intelligence gives artefacts the ability to display ethical behavior. For example, the ability to set goals and reason a variety of means to achieve them that result in certain ends always contains some ethical dimensions. Another may be the capacity to actively detect specific moments during instantiated behavior which have ethical dimensions – an ethical situation awareness. One may think this a radical departure from traditional artefacts, which gain their ethical dimensions from use patterns and from design purposes. However, as AI systems are designed and used as well, it is clear that they retain all ethical questions that relate to traditional artefacts. They also introduce new dimensions to the question of ethical design in proportion to their autonomy in goalsetting, planning, decision-making and implementation, among others. At the very moment an AI system can set goals and reason a course of action and implement it, it is engaging in ethical behavior. It is a different question whether the designer has explicitly considered the system from an ethical perspective or taken steps to ensure that ethical abilities are implemented as part of the AI system itself.

Taking the ethical stance towards AI design naturally cascades into problems of implementation and to sources of ethics (see [1], for example). These are problems which must be tackled or at least acknowledged simultaneously. Two key questions arise. The first is how to design and implement such abilities into AI systems. The second is where to look for the basis for ethical reasoning – where should "ethics modules" derive their contents and their basis for reasoning? This article offers a conceptual discussion on these issues and then lays out how cognitive mimetics [2-6] can be used in the context of AI ethics.

2 AI Ethics as a Design Problem

Ethics and the achievement of ethical behavior in AI systems is fundamentally a design problem. Design is about satisfying design goals, requirements, and constraints – and in this context specifically the requirements set by ethical standards for AI systems. Here, as elsewhere in design, the general task is to achieve a good fit between the form of the design and its context [7]. In this particular case, the "form" of the AI system should fit well to general ethical standards and practices given by the context. The context consists of the task, the task environment and the task- or domain-specific culture and practice, as well as the wider culture in which the AI system will be nested.

There are at least two senses in which ethics play a role in AI development: as extrinsic or intrinsic to the AI system. A good example of extrinsic ethics are provided by the guidelines set forth by the EU, the seven requirements for achieving 'trustworthy AI': human agency and oversight; robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness, societal and environmental well-being; and accountability. While these requirements can (and will) result in particular kinds of implementations, they are more focused on the use of AI and are directed at the designers as general obligations rather than specific instructions – they are, as it were, "outside" the technical system. By intrinsic we mean that the AI system itself has some capacity to evaluate its own behavior (or its outcomes) or external events from an ethical standpoint.

The question of values and ethics in information system design have been around for a while, a prominent example of which is Value Sensitive Design (VSD) [8-10]. VSD has spawned numerous articles over the decades. As a design framework for values, VSD has a tripartite structure consisting of conceptual, empirical, and technical investigations. While sound in principle, the framework has drawn also criticism, for example, Albrechtslund [11] has critiqued VSD for not sufficiently distinguishing between ethical design goals and non-ethical use patterns. For a response and further development of VSD see [12]. Van de Poel [13] has sought to fill the gap between VSD and implementation by way of a method for transforming values into design requirements. Van Wynsberghe and Robbins [14] call for a pragmatic approach to bring together ethicists and engineers in the lab to formulate values into technical systems. This is all sound in principle. However, not much has been written on the application of VSD in the context of AI [15], and it is possible the framework is not as such suited for the special problems of AI where the question of ethical information processing in the machine itself becomes central. The typical outputs of a VSD process are what we have called extrinsic ethics – general guidelines, constraints, and requirements.

What is crucial going forward, is to begin considering the ways in which human values, norms and ethics can be embedded into the systems, namely, how to implement moral reasoning *into* the AI system itself (steps towards this direction have been taken for example in [1]). This has been called machine morality or machine ethics [16, 17, 1]. This is closer to what we have called intrinsic ethics. The question for machine ethics is fundamentally about how machines could support or replace humans in performing ethical reasoning [18]. This makes the problem a special question within the general discourse of AI, but one that has a wide range of implications beyond the its own specific problems. In fact, ethical dilemmas and design problems have a surprising similarity: they are often open-ended and ill-structured. This means that in both, there are many acceptable solutions and in neither is there a routine process by which one can reach a solution. This is of course the primary source of trouble for AI systems – and the problems go far beyond ethics-specific questions. Thus, we have truly a wicked design problem at hand: the only thing we know perhaps is that universal formal ethical systems are surprisingly feeble for real-world reasoning and at any rate contested by philosophers. The tension is that given the operating principles of computers, those are (when formal, axiomatic and rules-based) the best suited for machine implementation. On the other hand, one might take the "bottom-up" approach [16] and seek to model in a neural network the patterns of behavior immanent in some context in a sub-symbolic fashion. However, if a machine (or an evaluating human for that matter) cannot justify and provide reasons for its behavior, can it be called ethical at all? The questions go deep into the foundations of AI and indeed push the envelope for AI development. The point is that ambitious designers and engineers should welcome, rather than shun as problematic, the challenge of intrinsic machine ethics. As the problem is far from being solved, as is the case for AI in general [19], there is also room in the discourse for fresh approaches and machine ethics provides as challenging a framework as any to advance these questions.

We take the primary purpose of AI to be to replace or support human information processing. We further take it to be the case that ethics are an instantiation of human information processing. Thus, in principle, this means that the system should somehow be able to evaluate situations and behaviors from an ethical standpoint (thus displaying a form of intelligence in this context). Typically, this would mean that it should not only be able to perform the objective task but engage in ethical information processing as well. The system should possess a functioning "ethics module".

If the machine is controlling the joining and disjoining of railway lines, and people happen to be strapped in uneven numbers on both paths the train is headed, it should be able to engage in a form of moral reasoning [20]. However, artificial toy problems such as the trolley problem easily box our thinking in. For AI design, it is important to understand *actual* ethical information processing in humans. As the trolley problem and its variations show, actual human ethical judgement is a complex affair which integrates many kinds of information processes and contents together against open-ended problems with many possible solutions. An AI system built around a single variable as the target of an ethical evaluation function may work for many cases but fail (from an ethical standpoint) for others because it has not mimicked [2-6] the actual information process in humans. For example, human beings judge based on emotions, or concepts like allowing vs. doing harm [20]. More importantly perhaps, examples like the trolley problem are artificially limited and designed to summon specific moral dilemmas. In real world situations (and even in toy problems), humans have the ability to restructure the problems beyond what the experimenter has in mind. In the trolley problem, for example, the choice to sacrifice oneself rather than harm others, is typically forbidden for artificial reasons, but in fact shows the highest moral virtue. Thus, accessing by empirical means (without artificial limitations) how humans restructure problems [21] can give crucial hints on how to build similar abilities into AI systems, and discover usable patterns for moral problem solving. The empirical route sketched in cognitive mimetics [2-6] provides implementation cues for both general abilities and specific contents and patterns [see 6]. The whole point of the trolley problem is that within its limitations there *is no* right answer, and in such unfortunate circumstances it is difficult to see how either humans or machines should be forced to consider ethical questions, as both making the choice and not-making it result in an unethical action in one sense or another. The right answer to the trolley problem is of course to stop the train from moving or perhaps to remove the people from the tracks. As a speculative example, the design answer to the problem should *not* be to have the machine calculate least number of victims (perhaps modulo age, health, etc.) and cause their death. We should simply build a remote control which can stop any train in its tracks within the needed timeframe. This is what ethical design thinking *should* be about when dealing with complex real-world problems that in reality admit to many different solutions. The focus for AI designers should perhaps be less on developing a moral calculus, and more on problem restructuring that requires none. From this perspective, a moral calculus may be more of a warning signal that problem restructuring is needed.

3 Cognitive Mimetics for Ethics

Cognitive Mimetics is an idea for a design method for intelligent systems introduced by Kujala and Saariluoma and elaborated over a series of papers [2-6]. Mimetic design means using a source in the natural or artificial worlds as an inspiration for technological solutions. In biomimetics one typically imitates the biological structures found in nature. However, in creating intelligent technologies designers can use existing organizational and individual information processes as the source of ideas. Designing intelligent systems by utilizing existing human information processes as the source of solutions we have termed 'cognitive mimetics' [2–6]. Cognitive mimetics differs from typical and established biomimetics as it has different source of mimicking: human shared and individual cognitive processes, as well as the mental contents, representations, and constraints that establish the boundaries and forms it takes. It analyses how people carry out intelligent tasks today and uses this information in designing novel technological solutions.

From the perspective of cognitive mimetics, ethics in action is fundamentally an instance of human information processing. Thus, the basic rationale of cognitive mimetics works here as well. The logical structure of mimetic design consists of three main parts [6]. To be an instance of mimetic design, there must be a source domain. The logical corollary to the source is the target domain. Furthermore, there is a process of interpretation or translation between the source and target domain, which can be called mimetic transfer. Implicit here is the designer who can extract and implement design-goal- relevant information from a source. Important to note is that the process of interpretation is observer-relative given that designers with different backgrounds and knowledge observe different aspects in the source [22]. Thus, as noted by Van Wynsberghe and Robbins [14] in the context of VBD, it is important to involve experts from different domains into the design process (ethicists and engineers). In cognitive mimetics, one would also include subject matter experts from the domain into which AI is being developed (whose actual ethical information processing is being supported or replaced) and ethicists, and then take both the content and the processes of their thinking into account in developing intelligent technology. The mimetic perspective is about finding out what makes the source an effective solution [6], which in the context of human thinking typically implicates concepts such as problem spaces [21], their construction and structure, heuristics, reasons, and mental representations in general. Here effectiveness can mean the mental representation's effectiveness as a solution to a moral dilemma. A very effective solution removes moral considerations completely or mitigates them significantly.

3.1 Tacit knowledge in ethics

Tacit knowledge is of special interest for cognitive mimetics [4]. Human information processing, ethical or otherwise, is typically grounded in a complex web of tacit knowledge. This forms a significant corollary to the mere "figure" of our thought readily available for introspection and verbal reporting. Indeed, much of the norms that guide ordinary life recede into the background so long as no one violates them. Tacit knowledge is essentially non-codified, disembodied know-how whose take up is often informal [23]. This sort of knowledge comes in various grades, some outside the possibility of explication and others not. For instance, while it is very difficult if not impossible to describe and transfer the smell of coffee to someone who does not have the same sensory experience, guiding another person to make coffee is relatively straightforward even though the process is quite automatic for most people in ordinary circumstances. Of course, in the latter example one can imagine how much tacit knowledge and skills is required for someone to be able to follow the instructions, and in this sense tacit knowledge provides a key which opens the pandora's box of how much is taken for granted in ordinary human life. The ability to follow instructions typically presupposes a vast amount of knowledge and skills in the tacit domain. This is very important for AI development, where such things cannot be taken for granted but must typically be programmed into the system [24]. Of course, recent developments in machine learning may in some sense lessen the burden of programming minutiae into the system, but equally clearly simply assuming the form of human behavior is not ethical in any real sense, and in fact one can imagine the opposite to be the case. A middle ground is likely to be necessary, least burdensome and most fruitful.

Ethical questions are likely to follow this tacit quality. With respect to previous work on cognitive mimetics, in the ontological schema [5] ethical questions are most likely to occur in the space of reasons (for action). It should be noted that action can be understood also as non-action and still remain intentional. These typically answer the question "why?" or "why not?". In the seafaring context studied in previous research [4-5], ethics are in fact absolutely central. In the research we did not even conceptualize the reasons for avoiding the ships on collision course from an ethical standpoint, although it is obviously at root an ethical and moral reason for behavior. Thus, in future research we can assume that much of ethical reasoning will follow this pattern and have deep tacit dimensions. It is also likely that we can't know these simply from our armchairs but must investigate actual activities.

"Why?" is the central guiding question for mimetic design. In terms of ethical information processing, this would typically lead through a succession of layers of meaning and lands at the ground of axiomatic assumptions. This type of investigation reveals the structure of ordinary knowledge, much of it tacit, and provides a central insight into the ontology (knowledge structure and contents) of the domain under investigation and its culture and norms.

3.2 Action ontologies for ethics

Ontologies are central for AI [25]. As all information systems – human, AI, or other – traffic in knowledge, it is important to be able to describe and organize domain-relevant information. Typically, some parts of the ontology are task-specific and others general [25]. What is relevant and sufficient in a particular case must be determined case-by-case. In all cases, it is necessary to enter into empirical investigation. Thus, the necessary ontology for ship steering [5] is probably different from the ontology of governmental complaint handling. And most certainly the contents will be different. Nevertheless, in all cases we can ask the question "What does the system know?" [25]. For cognitive mimetics this question is posed for the human operators (and other subject-matter experts) and later transformed into "what does the system *need* to know?" for AI systems.

In [5], we outlined a simple ontological structure for ship steering consisting of *observing, handling, and reasons.* It seems likely that with relatively minor adjustments this general ontology can work in many other contexts. Observation and interpretation are likely to be important in all AI contexts. Same applies to handling when conceptualized as behavior and action. By analogy these can be thought of as possible moves in a game. Reasons are perhaps the most crucial when it comes to ethics. Here, ethics can be thought of as a property of actions, as a constraint on actions, or as a goal for actions. They also provide a partial basis for what is interpreted in observation. Implicitly ethics can structure the whole ontology. For example, in autonomous cars a salient example is the interpretation of the visual scene of a ball bouncing from behind a car. Here the technical challenge is how to interpret the bouncing ball in terms of something that is not immediately seen (a child playing with the ball with the likely intention of running after it). The dilemma is of course completely ethical and has to do with reasoning beyond the immediately visible environment. Thus, the problems posed by ethics can often cascade into challenging technical questions that are in part answered by knowledge systems or ontologies. Here causal reasoning is implied in terms of a general ability for the AI system [19], but equally necessary are knowledge contents that give meaning to such interpretations.

3.3 Problem restructuring

Problem spaces [21, 26] offer another perspective on the ethics problem in AI. A problem space is a mental construct in which the human operates by what Newell and Simon called 'heuristic search'. The problem space is a representation of the possible solutions that a solver might consider for a given problem [26]. It is specified by the mental representation of the problem, the goal to accomplished and a set of actions (or operators) [26]. When applied the result is a solution path or a trajectory through the problem space.

Moral reasoning can be conceptualized in many senses: as goals, as constraints, or as problems. For example, I may have a more-or-less amoral goal (get ice cream), but the means by which I go about obtaining ice cream are constrained by ethical standards (I will buy it, not steal for example). One might say I have then a general high-level goal of ethical virtue (do not steal) which constrains some parts of my normal day to day life. Whether I know it or not (see previous on tacit knowledge and ontologies), ethical demands have thus operationally narrowed my problem space in the ice cream problem.

Let us consider moral dilemmas vis a vis problem spaces and problem space restructuring. Two perspectives are crucial. The first is to discover empirically in specified task environments the problem spaces individuals construct, as well as the paths they take to achieve their goals in an ethical fashion. What gave Newell and Simon [21] trouble, was the discovery of "significant interpersonal differences in processing", which made it difficult to describe problem-solving by a single computer model [26]. However, in moral dilemmas, as illustrated by the ability of people to reconstruct the problem spaces, this carries a clear benefit: by non-restrictive experimental settings, we may discover problem restructuring in action and thus different ethical problem-solving methods and problem representations for AI implementation. The single model paradigm should thus be shifted into another level and discover (as a likely distant goal for AI) how problem restructuring can take place in natural and artificial systems. The lesson for AI development is to focus less on myopic single-model or ethical paradigm solutions to moral dilemmas, and rather go beyond to the different ways in which actual humans restructure problems. This shifts the design focus away from moral calculus based on, for example, factors and weights, structured perhaps around a moral system. A moral calculus may be important, but one can argue that problem restructuring is both a more pragmatic approach by way of the contents and problem-solving methods it can provide, and also a deep long-term challenge in terms of an *ability* for the AI system.

3.4 Cognitive abilities and mental contents

Cognitive mimetics makes a conceptual distinction between cognitive abilities and contents [6]. Abilities here are general and necessary for all cognitive acts. Contents are domain-specific and learned or picked up by cognitive subjects over time. Take the example of problem restructuring. The ability of human subjects to restructure problem spaces is an ability we have. But the form and contents of our thinking that results in a problem space or the problem space itself are mental contents. We have suggested [6] the latter as a pragmatic starting point for AI design, since it is not clear whether the *ability* is a at present a realistic goal for actual current needs. The underlying idea is that mimetic design is based on the concept of multiple realizability, and in AI the common ground between computers and humans is to be found at the information level [2-3]. This information level can be realized to some extent in both platforms, even though their qualitative nature is very different.

As there seem to be fundamental limits to computational processes [27-28], the mimetic design process is about finding the common ground and representation form and accepting a large difference in implementation. Thus, cognitive mimetics begins with domain- and task-specific mental contents, ontologies, and thought processes rather than general abilities. General abilities, like problem restructuring, become implicated but the approach of cognitive mimetics leaves plenty of room for implementation strategies. Cognitive mimetics highlights the importance of domain-specific mental contents as starting point for AI development. Foundational work on AI abilities can then be fitted to the needs of this common ground.The approach in cognitive mimetics is no more about the material basis of information processes than a Turing Machine is about brain cells *or* transistors. The level at which cognitive mimetics approaches design problems in AI is the level of information processing and contents. The ethics problem in particular highlights that the correct level at which to look for ethics is not to be found in neurons, transistors, or even algorithms as such but at a higher level of abstraction.

4 Conclusion

The purpose of this article was to conceptualize ethical action and thought via the lens of human information processing and explore how some of the ideas in cognitive mimetics could be used in the context of AI ethics. Although no empirical work has yet been carried out, it seems clear that some of the key viewpoints in cognitive mimetics can be used to formulate ideas for designing ethics into AI. Tacit knowledge, action ontologies and problem structuring in ethical thinking are some clear examples of how concepts from the sciences around human information processing can be reflected onto the problem of ethics in AI.

Ethics and ethical information processing is an important problem for AI development. It is both important for obvious reasons, but also challenging in its' implications and should thus be welcomed by designers and developers as a major technical challenge. Cognitive mimetics may be useful in plotting out and providing contents for solutions in the intersection between man and machine and laying out necessary abilities that must be implemented for ethical information processing in machines to become a reality.

Acknowledgements

This work is funded by the Finnish Academy through the ETAIROS (https://etairos.fi/en/front-page/) project.

References

1. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE (2019).
2. Kujala, T., Saariluoma, P.: Cognitive mimetics for designing intelligent technologies. Adv. Hum. Comput. Interact. (2018).
3. Saariluoma, P., Kujala, T., Karvonen, A., Ahonen M.: Cognitive mimetics - main ideas. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), pp. 202–206. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2018).
4. Saariluoma, P., Karvonen, A., Wahlström, M., Happonen, K., Puustinen, R., Kujala, T.: Challenge of tacit knowledge in acquiring information in cognitive mimetics. In: Karwowski, W., Ahram, T. (eds.) Intelligent Human Systems Integration 2019. IHSI 2019. Advances in Intelligent Systems and Computing, vol. 903. Springer (2019)
5. Saariluoma, P., Wahlström, M., Kujala, T., Puustinen, R., Karvonen, A., & Happonen, K. An Ontology for Cognitive Mimetics. In 2018 International Conference on Computational Science and Computational Intelligence (CSCI)(pp. 1188-1192). IEEE (2018).
6. Karvonen, A., Kujala, T., & Saariluoma, P. Types of Mimetics for the Design of Intelligent Technologies. In International Conference on Human Systems Engineering and Design: Future Trends and Applications (pp. 40-46). Springer, Cham (2019).
7. Alexander, C. Notes on the Synthesis of Form (Vol. 5). Harvard University Press (1964).
8. Friedman, B., & Kahn Jr, P. H. Human values, ethics, and design. In The human-computer interaction handbook (pp. 1209-1233). CRC Press (2002).
9. Friedman, B. Value-sensitive design. interactions, 3(6), 16-23 (1996).
10. Friedman, B., Kahn, P. H., & Borning, A. Value sensitive design and information systems. The handbook of information and computer ethics, 69-101 (2008).
11. Albrechtslund, A. Ethics and technology design. Ethics and information technology, 9(1), 63-72 (2007).
12. Borning, A., & Muller, M. Next steps for value sensitive design. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1125-1134). ACM (2012).
13. Van de Poel, I. Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process (pp. 253-266). Springer, Dordrecht (2013).
14. Van Wynsberghe, A., & Robbins, S. Ethicist as Designer: a pragmatic approach to ethics in the lab. Science and engineering ethics, 20(4), 947-961 (2014).
15. Umbrello, S. Beneficial artificial intelligence coordination by means of a value sensitive design approach. Big Data and Cognitive Computing, 3(1), 5 (2019).
16. W. Wallach and C. Allen. Moral Machines: Teaching Robots Right from Wrong. New York: Oxford University Press (2008).
17. Anderson, M. & Anderson, S. L. Machine ethics. Cambridge University Press, New York (2011).

18. McLaren, B. M. Computational models of ethical reasoning: Challenges, initial steps, and future directions. IEEE intelligent systems, 21(4), 29-37 (2006).
19. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. Building machines that learn and think like people. Behavioral and Brain Sciences 40 (2017).
20. Woollard, Fiona and Howard-Snyder, Frances, "Doing vs. Allowing Harm", The Stanford Encyclopedia of Philosophy Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/doing-allowing/>. (Winter 2016 Edition).
21. Newell, A., & Simon, H. A. Human problem solving (Vol. 104, No. 9). Prentice-hall, Englewood Cliffs, NJ (1972).
22. Floridi, L.: The logic of design as a conceptual logic of information. Mind. Mach. 27(3), 495–519 (2017).
23. Howells, J. Tacit knowledge. Technology analysis & strategic management, 8(2), 91-106 (1996).
24. Dennett, D. C. Cognitive Wheels: the Frame Problem of AI. In M. A. Boden, *The Philosophy of Artificial Intelligence* (pp. 147-170). Oxford University Press, New York (1990).
25. Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. What are ontologies, and why do we need them?. IEEE Intelligent Systems and their applications, 14(1), 20-26 (1999).
26. Ohlsson, S. The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. The Journal of Problem Solving, 5(1), 7. (2012).
27. Saariluoma, P., & Rauterberg, M. Turing test does not work in theory but in practice. In Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 433). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2015).
28. Saariluoma, P., & Rauterberg, M. Turing's error-revised. International Journal of Philosophy Study, 4. (2016).