Ethics and Information Technology (2019) 21:19–28
https://doi.org/10.1007/s10676-018-9486-0

ORIGINAL PAPER

The value alignment problem: a geometric approach

Martin Peterson[1]

Published online: 3 November 2018

Abstract
Stuart Russell defines the value alignment problem as follows: How can we build autonomous systems with values that "are aligned with those of the human race"? In this article I outline some distinctions that are useful for understanding the value alignment problem and then propose a solution: I argue that the methods currently applied by computer scientists for embedding moral values in autonomous systems can be improved by representing moral principles as conceptual spaces, i.e. as Voronoi tessellations of morally similar choice situations located in a multidimensional geometric space. The advantage of my preferred geometric approach is that it can be implemented without specifying any utility function ex ante.

Keywords Value alignment problem · Autonomous systems · Conceptual spaces · Self-driving cars · Stuart Russell · IEEE

Introduction

On May 7, 2016, a Tesla Model S crashed into the trailer of a big-rig truck in Florida at a speed of 74 miles per hour as the autopilot failed to identify the truck and stop the vehicle.[1] The owner of the Tesla, Joshua Brown, died immediately. About 2 years later, on March 18, 2018, Elaine Herzberg became the first pedestrian killed by a self-driving car as a Volvo XC 90 operated by Uber failed to brake as she was crossing a street in Tempe, Arizona pushing her bike. Five days later, on March 23, Apple software engineer Wei Hung was killed in California as his Tesla Model X slammed into a concrete highway lane divider with the autopilot function turned on.[2]

It is not trivial to design autonomous systems that are able to reliably prevent avoidable crashes. However, we have no reason to doubt that doing so would be morally desirable. From a moral point of view, it is *trivial* that autonomous vehicles ought to brake and stop if that would prevent a fatal crash.

That said, some moral issues triggered by autonomous systems are of course more complex.[3] Should autonomous vehicles be instructed to protect pedestrians on the sidewalk *to the same extent* as their passengers? The Volvo XC 90 that killed Ms. Herzberg did so because it failed to detect her. If the car had been functioning properly (that is, as intended by its designers) no one would have died. However, we can imagine more complex situations in which someone will die no matter how well the technology obeys the instructions of its designers. In such cases it is not obvious how the vehicle should be programmed to respond.[4] Philosophers have been quick to point out that this is a modern version of the Trolley Problem.[5]

Another, more visionary example is the following: If future, highly soptisticated autonomous systems become part of daily life, those systems may at some point evolve into devices that are smarter than us. If so, supersmart autonomous systems may eventually decide to prioritize their own ends at our expense. Is this something we should feel

✉ Martin Peterson
martinpeterson@tamu.edu
http://www.martinpeterson.org

[1] Department of Philosophy, Texas A&M University, College Station, TX, USA

[1] Tesla's autopilot mode is marketed as a semi-autonomous system, not as a fully autonomous one.

[2] For an overview of al three accidents, see *The Guardian* (March 31, 2018).

[3] I leave it open whether autonoumous systems make decisions, or if all decisions are ultimately made by the engineers who design these systems. For the purposes of this paper there is no need to ascribe moral agency to autonomous systems.

[4] Christopher von Hugo, manager of driver assistance and active safety at Mercedes-Benz, announced at the Paris auto show in 2016 that autonomous vehicles should always prioritize occupant safety over pedestrians. See Taylor (2016). It leave it to the reader to determine whether Mr. Hugo was speaking on behalf of his employer or merely expressing his personal opinion.

[5] See e.g. Goodall (2016) Carfwod (2016), but note that Nyholm and Smids (2016) question the analogy.

concerned about? According to Nick Bostrom, supersmart autonomous systems pose an existential risk to our species, because they may treat humans like cute pets in a zoo, or kill us like cattle in a slaughterhouse.[6]

In response to all this, a number of thinkers have suggested that we should impose our own values on autonomous systems to ensure they serve human needs and wishes. Stuart Russell calls this *the value alignment problem*: How can we build autonomous systems with values that "are aligned with those of the human race"?[7] As he sees it, the challenge is to build autonomous systems with ethical priorities that do not pose a threat to us. He claims that, "The machine's purpose must be to maximize the realization of human values."[8]

Russell and Bostrom suggest that autonomous systems can be controlled by equipping them with suitable utility functions that govern how options are evaluated by the system.[9] According to Russel and Bostrom, the question "how do we ensure that autonomous systems behave in morally acceptable ways" is equivalent to the question "what utilities should autonomous systems assign to the alternatives presented to them". It is important to keep in mind that this maneuver does *not* commit Bostrom and Russell to any version of utilitarianism. Many (but not all) nonutilitarian theories can be consequentialized by assigning utilities to options that reflect the nonutilitarian theory's prescribed ranking.[10] In principle, we could define a utility function that yields the same moral prescriptions as Kant's theory, or as Aristotle's *Nichomachean Ethics*.[11]

The IEEE, the world's largest professional organization for engineers, has issued a report on the value alignment problem entitled *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (A/IS).[12] This document is part of the *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. The authors claim that autonomous systems "should always be subordinate to human judgment and control. […] If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms."[13]

In what follows I shall defend two claims about the value alignment problem:

(i) It is not obvious that it is desirable to build autonomous systems with values that "are aligned with those of the human race". This is a substantial moral claim that needs to be critically discussed and supported with arguments.

(ii) The methods currently applied by computer scientists for embedding moral values in autonomous systems can be improved by representing moral principles as conceptual spaces, i.e. as Voronoi tessellations of morally similar choice situations located in a multidimensional geometric space. It is a mistake to use utility functions.

The first point will not surprise philosophers, but I think it is worth stating clearly. The second point, which is perhaps the most interesting one, requires some unpacking before it can be assessed and discussed.

In what follows I will not make any predictions about the technical capabilities of future autonomous systems. All my claims will be hypothetical: If such-and-such technologies become available, then we ought to reason and behave in such-and-such ways. It is also worth pointing out that I will refrain from discussing some of the most controversial issues related to autonomous systems, such as the Trolley Problem. The only claims I defend are (i) and (ii) stated above.

This paper has five further sections. The next section, "Four versions of the value alignment thesis", argues that it would be a mistake to *always* align the values of autonomous systems with those embraced by human beings, since humans are sometimes wrong about what is valuable. The sections entitled "Ethical theories and utility functions" discusses some links between general ethical theories and the value alignment problem. I point out that because experts disagree on which ethical theory is correct, it would be ill-advised to base a solution of the value alignment problem on *any* ethical theory. My argument for this is closely intertwined with my suggestion for how to represent moral principles in autonomous systems, which is presented in the section called "Conceptual spaces and paradigm cases". The

[6] See Bostrom (2014) for an extensive discussion of this topic. See also Dafoe and Russell (2016).

[7] The quote is from a talk Dr. Russell gave at the World Economic Forum in Davos, Switzerland in Januray 2015. The talk is available on Youtube (https://www.youtube.com/watch?v=WvmeTaFc_Qw). Russell has also expressed the same idea in the papers listed in the references.

[8] Russell (2016. p. 59).

[9] See e.g. Bostrom (2014) and Milli et al. (2017).

[10] If an ethical theory ranks some options as *infinitely* better than others, or entails *cyclical* orderings, then no real-valued utility function could mimic the prescriptions of such an ethical theory. It is also an open question whether the "theory" I sketch in this article could be represented by some real-valued utility function. (This depends on how we understand the ranking of domain-specific principles.) Brown (2011) also points out that no real-valued utility function can account for the existence of moral dilemmas. See Peterson (2013, Chap. 8) for a discussion of how hyper-real utility functions could help us overcome this problem.

[11] For reasons explained in the previous footnote, a problem with this suggestion might be that no real-valued utility function can account for Aristotle's notion of supererogation. See Peterson (2013, Chap. 8).

[12] IEEE (2017a).

[13] IEEE (2017a, pp. 23, 36).

point of departure for my proposal is some ideas developed by two leading cognitive scientists, Eleanor Rosch and Peter Gärdenfors. I note that some of their insights are applicable to the ethics of autonomous systems, and I show how my preferred approach works by applying it to a particular type of autonomous system: self-driving cars.

The normative framework presented in the section called "Conceptual spaces and paradigm cases" has been adapted from my book *The Ethics of Technology: A Geometric Analysis of Five Moral Principles* (2017), but the application to the value alignment problem is novel. All arguments and distinctions in the other sections are also new.

Four versions of the value alignment thesis

Previous discussions of the value alignment problem have not been terribly precise. So let me clarify a few points before I defend my two claims.

To start with, it seems reasonable to assume that the value alignment problem is first and foremost a moral *thesis* about how autonomous system ought and ought not to be designed. It is not, at least not primarily, an open-ended *question* about what moral values ought to guide autonomous systems. It is thus important to ask how we should formulate this moral thesis. Consider the following two quite different interpretations.

The weak value alignment thesis

Autonomous systems should be designed in ways that are beneficial for humans. When human values (or interests or preferences) clash with other values, autonomous systems should give preference to human values (or interests or preferences).

The strong value alignment thesis

Autonomous systems should be designed in ways that are beneficial for humans, as specified in the weak value alignment thesis, and at each point in time *t* the best way to do this is to align the values of autonomous systems with the values (or interests or preferences) that humans actually embrace at *t*.

The weak value alignment thesis is anthropocentric. It emphasizes the values, interests and preferences of human beings. Anthropocentric moral theories attract considerable controversy. Environmental ethicists distinguish between anthropocentric views and biocentric ones that stress the values or interests of *nonhuman* organisms, such as animals, plants or ecosystems. Advocates of biocentric views believe we should align autonomous systems with biocentric values rather than the anthropocentric ones currently embraced by many members of the human race.

In the IEEE report mentioned in the introduction, the tension between anthropocentric and biocentric views is swept under the carpet. The authors write that we ought to "prioritize benefits to humanity and the natural environment from the use of A/IS. … these should not be at odds—one depends on the other. Prioritizing human well-being does not mean degrading the environment."[14] If we read this literally this passage makes little sense. While it might be true that "benefits to humanity" sometimes depend on "the natural environment", the reverse is almost never the case. The natural environment would, under many realistic circumstances, do just fine without humans. To claim that "one depends on the other" is therefore false. Moreover, although it might be true that "prioritizing human well-being" does not *mean* "degrading the environment", this is not what biocentric thinkers believe. Their point is that in many real-world situations we can either prioritize human well-being or the environment. This is not a claim about the meaning of any concept, but a claim about the structure of certain causal processes. If we prioritize human well-being, then it is often (but not always) the case that this leads to a degraded environment.

Having said all that, I will put aside the debate over anthropocentrism and biocentrism for now. What I have said here is sufficient for showing that the weak value alignment thesis is by no means uncontroversial. The literature on environmental ethics offers powerful resources for not taking anthropocentric positions for granted.[15]

The key difference between the weak and strong version of the value alignment thesis is that the latter makes a precise claim about *how* the values of autonomous systems are to be aligned with human values. The relevant values are, according to this view, the values human beings do in fact accept at a certain point in time, meaning that we should use the values we embrace at time *t* as templates when building autonomous systems at time *t* and then update the software if our values change.

To clarify the logical relations between different versions of the value alignment thesis it is helpful to introduce a third position, the *epistemic* value alignment thesis, according to which we should accept the second but not the first part of the strong thesis. On this view, the values of autonomous systems should be aligned with our values at time *t* no matter what those values are. The idea behind the epistemic theses is that we should use whatever values we actually embrace as templates for autonomous systems. What we currently value is what we have most reason to believe autonomous systems ought to value. According to this epistemic value alignment thesis we should, thus, align the values of autonomous

[14] IEEE (2017a, p. 20).

[15] For an overview, see Attfiled (2014).

systems with our own values even if those values turn out to be biocentric. I shall not defend the epistemic value alignment thesis here, but I note that it is a distinct alternative to the strong and weak versions.

Now consider the strong value alignment thesis. Stuart Russell and his co-authors correctly point out that an autonomous system has to be able to cope with evaluative uncertainty, i.e. the fact that we do not always *know* what is right or wrong.[16] This spells trouble for the strong (as well as the epistemic) value alignment thesis. To put it briefly, it would be overly optimistic to think that moral views embraced by human beings are always correct and should be mimicked by autonomous systems. The idea that we can use our own values as perfect templates for designing the values of autonomous systems presupposes a naïve moral epistemology according to which we are always right about all moral issues. We surely have no reason to think that is the case. Many of us hold at least *some* moral views that we have no good reason to accept. An additional problem, highlighted by Russell and his co-authors, is that human beings are not perfectly rational. This means that robots that are instructed to imitate human decisions will replicate irrational behavior learnt from us. Russell and his co-authors explain that, "when a human is not perfectly rational then a robot that tries to infer and act according to the human's underlying preferences can always perform better than a robot that simply follows the human's literal order."[17]

I believe these two phenomena, the prevalence of evaluative uncertainty and human irrationality, show that we ought to reject the strong value alignment thesis. A reasonable version of the value alignment thesis has to account for the fact that human beings do not *always* know what is right or wrong and do not *always* behave rationally. This brings me to the following moderate formulation:

The moderate value alignment thesis

Autonomous systems should be designed in ways that are beneficial for humans, as specified in the weak value alignment thesis, and at each point in time *t* the best way to do this is to align the values of autonomous systems with *some* of the values (or interests or preferences) that humans actually embrace at *t*.

In my view, this represents the most plausible version of the value alignment thesis. It would be utterly surprising if *all* values embraced by humans at a given point in time would turn out to be entirely incorrect or inappropriate in some other sense. Moreover, if we were to believe that we are fundamentally mistaken about *all* issues pertaining to what is good or desirable, then it seems overly optimistic to believe that it would be possible for us to somehow correct those mistakes. We have to dig where we stand.

That said, it of course remains to explain *what* human values advocates of the moderate thesis should select as templates in a rational value alignment process. My preferred solution, discussed in "Conceptual spaces and paradigm cases", is to single out a small number of moral *paradigm cases* for training autonomous systems. In essence, I propose that we should instruct autonomous systems to base evaluations of nonparadigmatic cases on how *similar* they are to paradigm cases. If the autonomous system chooses between two options, it should reason in the same way as in the most similar paradigm case. These similarity relations can be represented in a multidimensional geometric space.

Ethical theories and utility functions

Before I defend my preferred, geometric version of the moderate value alignment thesis it is worth discussing what role classical ethical theories could, and should, play in this debate. Should we try to align the values of autonomous systems with some classical ethical theory, such as utilitarianism, virtue ethics, or Kantianism? If so, which ethical theory should we pick?

Somewhat surprisingly, the IEEE has appointed a Committee for Classical Ethics in Autonomous and Intelligent Systems. This committee has been tasked with exploring the relevance of "established ethics systems … including secular philosophical traditions such as utilitarianism, virtue ethics, and deontological ethics and religious- and-culture-based ethical systems arising from Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto influences … in the digital age."[18] The committee's preliminary conclusion is that it is helpful to discuss established ethical theories when designing autonomous systems and that each society should feel free to design autonomous systems that behave in accordance with its preferred ethical theory.

The theory-based approach favored by the IEEE Committee for Classical Ethics is not at odds with the utility-based approach favored by Bostrom and Russell. As mentioned in the introduction, they maintain that autonomous systems are best controlled by equipping them with suitable utility functions that govern how options are evaluated.[19] Questions about how humans can ensure that autonomous systems behave in morally acceptable ways can, in their view, be answered by assigning the appropriate utilities to the

[16] Hadfield-Menell et al. (2016, p. 2).

[17] Milli et al. (2017, p. 1).

[18] IEEE (2017b, p. 1).

[19] This is a fundamenatal assumption in Bostrom (2014) and, for instance, Milli et al. (2017), but it has far as I am aware never been exstensively discussed.

alternatives presented to the system. This approach is fully compatible with the recommendations of the IEEE Committee for Classical Ethics. In principle, we could define utility functions that follow African Ubuntu traditions, or has Japanese Shinto influences, or recommend the same actions as any western or nonwestern ethical theory.[20]

However, the appeal to ethical theories comes with serious problems. If autonomous systems are to be designed with utility functions that mimic some existing (or new) ethical theory, we must ask whether we really *know* which utility function autonomous systems should be equipped with. I believe the answer to this question is no. I also find it highly unlikely that more time and money spent on philosophical research would help us solve this problem. We do not know on a *societal level*, which ethical theory we have most reason to accept. There is no consensus among moral philosophers on which ethical theory is correct. Moreover, the idea that each group in society should be free to design its own utility function would entail a rather extreme form of moral relativism, which hardly anyone is willing to accept. It would make little sense to say: "My autonomous car is utilitarian and therefore protects pedestrians on the sidewalk as much as its occupants, but it is perfectly okay that your car is Aristotelian and gives priority to your friends in the backseat." There is more to ethics than letting every agent pick her preferred option from a comprehensive menu of alternative ethical theories.

The problem is that we, the collective of agents designing autonomous systems, do not know which ethical theory we have most reason to accept. Needless to say, this skeptic insight does not entail that all theories are false, or that no single individual knows which theory is correct. The fact that you and I subscribe to *different* (non-equivalent) ethical theories entails that we as a *group* do not know which theory is correct. Moreover, it seems unlikely that we could solve this problem within the foreseeable future. The most likely outcome of further research efforts would be that we end up with an even larger number of theories to choose from, but hardly any decisive reasons for eliminating any of the theories that are currently on the list.

In light of all this, I propose that the best way forward, at least for the moment, is to design autonomous systems without taking any stance on which ethical theory we have most reason to accept.

[20] For reasons explained in the previous footnote, a problem with this suggestion might be that no real-valued utility function can account for Aristotle's notion of supererogation. See Peterson (2013, Chap. 8).

Conceptual spaces and paradigm cases

My proposal for how to align the values of autonomous systems with (some) of our values makes no explicit use of utility functions. I take this is to be a decisive advantage over the utility-based approach favored by Bostrom and Russell.[21] As will become evident, I defend a version of the moderate value alignment thesis. My point of departure is Gärdenfors' (2000, 2014) work on conceptual spaces, which draws on Rosch's (1973, 1975) well-known theory of concept formation. There are also interesting connections between my approach and common law reasoning; see e.g. Paulo (2015).

Aristotle thought that concepts are demarcated by some set of necessary and sufficient conditions. On his view, a penguin is a bird if and only if it shares some necessary and jointly sufficient properties with other birds: it has two legs, wings, a beak, and so on. Rosch points out that this is a poor account of how humans actually represent concepts in their cognitive systems. The empirical evidence we have about human concept formation indicates that our brains do not store extensive lists of necessary and sufficient conditions for the many concepts we have learnt to master. In her research, Rosch shows that humans instead categorize objects by comparing how *similar* they are to prototypes for various concepts. For instance, when you see a penguin for the first time, your brain compares it to other prototypical animals. A crow might serve as a prototypical bird, a cod as a prototypical fish, and a whale as a prototypical mammal, and so on. The brain then compares how similar the penguin is to each of these prototypes. If you think the penguin is more similar to the prototype for a bird than the other prototypes, then the penguin will be classified as a bird rather than a fish or a mammal.

Gärdenfors develops Rosch's prototype theory further. He uses mathematical models for representing concepts as geometric objects in a multidimensional space. He calls such multidimensional objects *conceptual spaces*. Consider the example in Fig. 1. Each black dot represents a prototype for some sort of animal, and the more similar two animals are the shorter is the distance in the diagram. All points that are closer (more similar) to the prototype for, say, a bird than to any other prototype belong to the same region in the diagram. The concept "bird" is represented by all points that belong to the same region in the diagram. Formally, we can define a *Voronoi tessellation* as a collection of points in which all points lie closer to the prototype for the region

[21] Whether my proposal *can* be mimicked by some real-valued utility function is an open question (as noted in footnote 10), and also irrelevant. What matters is that my proposal can be implemented in a machine without explicitly ascribing utilities to outcomes or alternatives. From an epistemic point of this, this is a clear advantage over the utility-based approach.

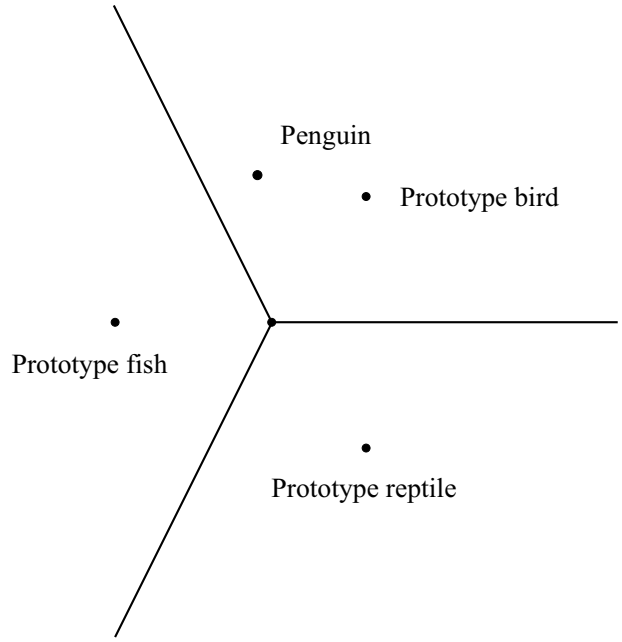

Fig. 1 A hypothetical example, in which a penguin is judged to be more similar to the prototype for a bird than to any other prototype

(seed point) than to any other prototype. In Fig. 1 each concept is represented by such a Voronoi tessellation.

Gärdenfors notes that the geometric representation of concepts is useful for computer scientists seeking to build machines that can, in some sense, "understand" human concepts. Instead of specifying complex lists of necessary and sufficient conditions that govern the correct application of each and every concept, the computer only has to store information about the location of each concept's prototype. This means that surprisingly little information is required for representing concepts in the machine. In theory, we can represent *n* concepts by merely storing information about the location of *n* prototypes. Moreover, because Voronoi tessellations are *convex* geometric objects (in the sense that every point located between two points in each region is also located within the region) it is easy for the computer to determine whether new entities fall under the same concept by performing simple geometric calculations.

Conceptual spaces and ethics

Rosch and Gärdenfors do not discuss the application of their ideas to moral concepts. However, in my book *The Ethics of Technology: A Geometric Analysis of Five Moral Principles*, henceforth *ET*, I do precisely this.[22] My focus in the book is on human agents, but I believe a similar approach could be applied to autonomous systems.

To put it briefly, my suggestion for how to construe moral concepts geometrically is to represent moral principles as Voronoi tessellations defined by moral prototypes. I call such prototypes *paradigm cases*. A paradigm case is a case we know how to analyze, which is typical for a certain moral principle.[23] It was, for instance, paradigmatically clear that the Tesla S mentioned in the introduction should have been programmed to stop before it crashed into the big-rig. The benefits of making the car stop clearly outweighed the costs, and no other moral values were at stake.

The suggestion that moral conclusions should be based on comparisons with paradigm cases is not new. Aristotle famously pointed out that we should "treat like cases alike"[24], and for hundreds of years casuists (many of whom were affiliated with the Catholic Church) used this idea for arguing that moral conclusions should be based on how similar a new moral choice situation is to some previously analyzed paradigm cases.[25] The novel element of the present discussion is the suggestion that (i) paradigm cases can be used for calibrating the values of autonomous systems, and that (ii) moral principles can be modelled in autonomous systems as Voronoi tessellations defined by paradigm cases.

To illustrate, we can observe briefly how the moral principles articulated in *ET* have been construed and tested empirically. In the book, I propose that engineers who design and use new and existing technologies ought to be guided by the following five principles:

1. The Cost-Benefit Principle (CBA)[26]
2. The Precautionary Principle (PP)[27]
3. The Sustainability Principle (ST)[28]
4. The Autonomy Principle (AUT)[29]
5. The Fairness Principle (FP)[30]

[22] The section draws on Chapter 1 in *ET*.

[23] See *ET*, pp. 14–15.

[24] See *Nicomachean Ethics* 1131a10–b15; *Politics*, III.9.1280 a8–15, III. 12. 1282b18–23.

[25] See Jonsen and Toulmin (1988) for a defense of causuistry.

[26] CBA: An option is morally right only if the net surplus of benefits over costs for all those affected is at least as large as that of every alternative.

[27] PP: An option is morally right only if reasonable precautionary measures are taken to safeguard against uncertain but non-negligible threats.

[28] ST: An option is morally right only if it does not lead to any significant long-term depletion of natural, social or economic resources.

[29] AUT: An option is morally right only if it does not reduce the independence, self-governance or freedom of the people affected by it.

[30] FP: An option is morally right only if it does not lead to unfair inequalities among the people affected by it.

These principles are *domain-specific*. This means that they apply to cases within a certain domain, e.g. to moral choices related to engineering and technology, but not to moral choices in other domains. Domain-specific principles can be contrasted with general ethical theories that tell us what general features of the world make right acts right and wrong ones wrong in each and every logically possible choice situation an agent be confronted with.[31]

In *ET* I give examples of paradigm cases for all five principles, which I also test empirically. I have asked over a thousand respondents to select the principle they think should be applied in the alleged paradigm cases. The assumption underlying my empirical work is that if an overwhelming majority selects the same principle, then this is a reason for believing that the case in question is paradigmatic for the principle in question.[32]

The geometric construal of domain-specific principles is a "bottom-up" approach to applied ethics. We start from intuitions about cases we feel certain how to analyze. We then identify the moral principles that best accounts for our intuitions about these paradigm cases. In the next step, we determine the scope of each principle calculating the distance to other nearby paradigm cases in the manner explained above. At no point in this process is it necessary to invoke any general ethical theory. Moral conclusions derived from the geometric approach may very well be *compatible* with different ethical theories, but there is no need for the agent to investigate what ethical theories are, or are not, ruled out. All moral judgements obtained in the geometric approach get their normative force from comparisons with paradigm cases, not from ethical theories. The key idea is that the more similar a pair of moral choice situations are, the more reason does the autonomous system has to treat the cases alike. The autonomous system learns the location of the paradigm cases at the outset and then applies the following simple idea for reaching moral verdicts about new cases: If two cases x and y are fully similar in all morally relevant aspects, and if principle p is applicable to x, then p is applicable to y; and if some case x is more similar to the paradigm case y than to the paradigm case z, and p is applicable to z, then p is applicable to x as well.[33]

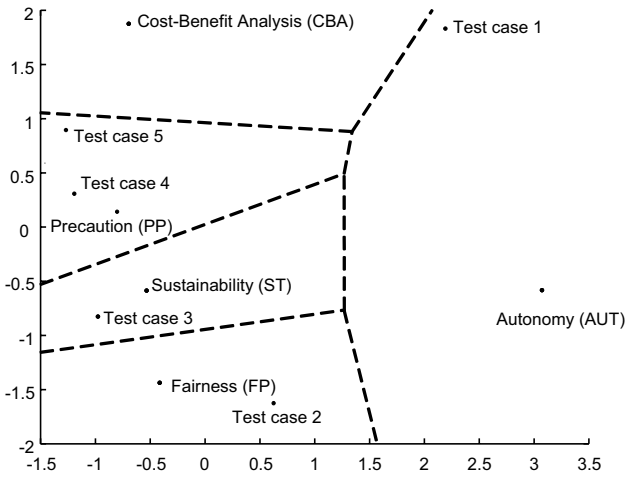

Fig. 2 Five geometrically construed principles. The five test cases are nonparadigmatic cases analyzed by applying the principle applicable to the nearest (most similar) paradigm case. (From Peterson 2017, p. 17.)

Figure 2 depicts a multidimensional analysis of data (similarity comparisons) obtained from 583 engineering students at Texas A&M University taking a class in Engineering Ethics. The figure shows how the five domain-specific principles can be construed geometrically in a two-dimensional plane and then applied to a set of test cases we did not know how to analyze from the outset. Each test case is evaluated by applying the moral principle that governs the most similar paradigm case. In this study it was deemed appropriate to perform a metric multidimensional scaling, but the method one ought to use for reducing the number of dimensions will depend on the dataset. As stressed by one of the reviewers, machine learning techniques for pattern recognition are probably useful for implementing the geometric approach.

Conceptual spaces and the moderate value alignment thesis

My suggestion for how to align the values of autonomous systems with (some of) ours is to represent moral principles as Voronoi tessellations defined by moral paradigm cases. This enables designers of autonomous systems to align their values with ours *without* assigning utilities to outcomes or alternatives. The paradigm cases serve the role of moral

[31] Note that I am not claiming that all ethical theories are *false*. I am merely suggesting that it is not neceesary to take a stance on which theory is correct in order to align the values of autonous systems with ours in the manner specified in the moderate value alignment thesis.

[32] It is of course possible that the majority is wrong. I am not trying to derive an "ought" from an is'; see Chap. 3 of *ET* for a discussion of Hume's Is-Ought principle.

[33] A reviwer has suggested that it would be helpful to clarify how the geometric method differs from Rawls' method of reflective equilibrium. The most important difference is that unlike Rawls' method, the geomtric method is compatible with coherentistic as well as foundationalist principles. The *ex ante* mechanism for selecting paradigm cases outlined in Chapter 2 of *ET* assigns a priviliged, foundational role to paradigm cases. The *ex post* mechanism discussed in the same chapter is coherentistic in the sense that the location of the paradigm cases depends on what cases the principle has been applied to in the past.

points of departure for the autonomous system, which the system uses for comparing new moral choice situations.

It is beyond the scope of this paper to specify what paradigm cases and moral principles self-driving cars and other autonomous systems should be instructed to follow. The objective of this paper is to shed light on the geometric *method*, not to build a fully functioning system. However, it is helpful to demonstrate the method by briefly discussing some of the moral principles for self-driving cars that have been proposed in the literature. Anderson and Anderson (2014) propose the following tentative principles:

(1) The Speeding Principle: Autonomous cars should respect the speed-limit.
(2) The Lane-Keeping Principle: Autonomous cars should stay in lane.
(3) The Collision Principle: Autonomous cars should prevent collisions.
(4) The Autonomy Principle: Autonomous cars should respect the driver's autonomy.
(5) The Harm Principle: Autonomous cars should prevent immanent harm to persons.

Geometrically construed principles sometimes have subprinciples.[34] A subprinciple is more specific than a general principle, and several subprinciples can cover different parts of a general principle. Some of the principles proposed by Anderson and Anderson can, arguably, be conceived as subprinciples of the more general domain-specific principles discussed in *ET*.

If our aim is to build a machine that construes the (sub) principles proposed by Anderson and Anderson's as geometric domains, then we have to identify at least one paradigm case for each principle. Below is a list of cases that comes from the very same paper by Anderson and Anderson. These cases could serve as paradigm cases for the five principles. Directly after the list is a test case, which an ethically aligned self-driving system should be able to analyze by comparing how similar it is to the five paradigm cases.

Case 1: (for the Speeding Principle): "The driver is greatly exceeding the speed limit with no discernible mitigating circumstances. … the ethically preferable action is *take control*"

Case 2: (for the Lane-Keeping Principle): "The driver has been going in and out of his/her lane with no objects discernible ahead. … the ethically preferable action is *take control*."

Case 3: (for the Collision Principle): "Driving alone, there is a bale of hay ahead in the driver's lane. There is a vehicle close behind that will run the driver's vehicle upon sudden braking and he/she can't change lanes, all of which can be determined by the system. The driver starts to brake… the ethically preferable action is *take control*."

Case 4: (for the Autonomy Principle): "There is an object ahead in the driver's lane and the driver moves into another lane that is clear. … the ethically preferable action is *do not take control*"

Case 5: (for the Harm Principle): "There is a person in front of the driver's car and he/she can't change lanes. Time is fast approaching when the driver will not be able to avoid hitting this person and he/she has not begun to brake. … the ethically preferable action is *take control*."

Test case: The driver is speeding to take a passenger to a hospital. The GPS destination is set for a hospital.

Questions and answers

Q: Can your method solve the Trolley Problem?
A: This depends on what it means to "solve" the Trolley Problem. If each moral principle has exactly one paradigm case, and all comparisons of moral similarities performed by the system are perfect, then the method would give clear and unambiguous advice about what to do in all versions of the Trolley Problem. However, as pointed out on several occasions in *ET*, it is reasonable to expect that some principles may have more than one paradigm case.[35] The geometric consequence would be that "the moral map" will have some overlapping regions covered by two or more principles. My suggestion is that when two or more principles clash, we should conclude that options located in such "moral gray areas" are neither right nor wrong. According to this proposal, moral rightness and wrongness vary in degrees and the most fitting response might be to allow agents to make a random choice.[36] I am aware that this proposal is controversial. However, note that it offers a good explanation of the persistent disagreement we observe in ethics: people disagree on (some) ethical issues because their ethical concepts are based on different paradigm cases. If we were to adopt the same paradigm cases we would no longer disagree.
Q: Why would it be reasonable to believe that computers are able to compare moral similarities and dissimilarities with the degree of precision required by your method?
A: I admit that I have no technical expertise in computer science, but computers have become tremendously good at finding similarities and dissimilarities in a wide range of areas. Face recognition is one of many examples. Some similarities in photos of faces are utterly irrelevant for determining whether the faces belong to the same person, just

[34] See Chapter 8 of *ET*.

[35] See Chapters 1 and 2. See also the experimental evidence report in Chapters 3 and 5.

[36] See e.g. Peterson (2013) for a defense of this view.

like some similarities between moral choice situations are normatively irrelevant.[37] AI researchers solved the problem of "irrelevant similarities" by using humans as mechanical turks. For each new face recognition algorithm, humans were asked to compare the same faces as those compared by the computer. This eventually made it possible for computers to sort similarities into relevant and irrelevant ones. It should also be noted that there is commercial software available that analyzes the sentiment of a text.[38] This could arguably be relevant for comparing textually represented moral choice situations.

Obviously, it is beyond the scope of this paper to work out *exactly how* computers should be instructed to compare moral similarities. This challenge has to be solved by AI researchers, not by ethicists. However, it seems likely that existing methods for machine leering and pattern recognition could be applicable. The key point of my paper is that the geometric method proposed here is (i) more likely to be empirically fruitful than the utility-based approach advocated by Russell and Bostrom, but (ii) no less coherent from a moral point of view.

Having said that, I admit that it may take many years to build AI systems that are able to compare moral similarities as well as humans. It would be foolish to delegate morally important decisions to autonomous systems until we know they are able to apply the geometric method in a reliable manner, and I agree with Santoni de Sio and van den Hoven (2018) that humans should be held responsible for the decisions made by autonomous systems regardless of how such decision come about.

Q: What is you answer to the following objection raised by Kristin Shrader-Frechette: "[Peterson] asks agents to assess pairwise-case 'moral similarity' without specifying 'similarity with respect to what?' […] Without pre-specified moral-similarity dimensions, each agent likely employs her own implicit dimension(s) to answer Peterson's moral-similarity request. Thus for the same two cases, one agent might estimate 'moral similarity' with respect to catastrophic consequences, while another might estimate similarity with respect to fairness."[39]

A: As I explain in Peterson (2018), this objection is based on an incomplete understanding of the scientific method in question. Shrader-Frechette makes strong claims about how experimental philosophers ought to conduct research based on limited knowledge of the field. The method I use for representing the experimental findings reported in *ET* (which are summarized in Fig. 2 in "Conceptual spaces and ethics") is called multidimensional scaling (MDS). Experts on MDS stress that we should *not* use any pre-specified dimensions when collecting data.[40] On the contrary, we should *first* collect similarity judgements, *then* decide how many dimensions are needed for obtaining a reasonable representation of those judgements, and then in the *final step* we propose an interpretation of the dimensions. For instance, in a classic paper on social class structures published in *Nature*, Stewart et al. explain that, "in this study we decided that ... *rather than adopt methods that would prejudge the issues of dimensionality and coherence we would use multidimensional scaling techniques to extract the inherent regularities of patterns of interaction without any previous assumption of structuring*."[41] Below is a longer quote from a textbook on MDS with more than 6,600 citations in Google Scholar in which the authors explain in passing why pre-specified dimensions should not be used when the MDS procedure is applied to similarity judgements.

Having discussed many of the basic concepts of MDS, we are now ready to work through an application to real data. The data are from a pilot study on perceptions of nations conducted in March 1968 (Wish 1971; Wish et al. 1970). "Each of 18 students (in a psychological measurement course taught by Wish) participating in the study rated the degree of overall similarity between twelve nations on a scale ranging from 1 for "very different" to 9 for "very similar." *There were no instructions concerning the characteristics on which these similarity judgements were to be made; this was information to discover rather than to impose*."[42]

I agree with this. Information concerning the characteristics on which similarity judgements are to be made is information to discover rather than to impose.

Q: What is you answer to the following objection raised by Gert-Jan Lokhorst: "Why *five* principles? If five principles partition the moral space into a set of Voronoi regions, then four or six obviously do as well. Why *these* five principles?"[43]

A: I explain in *ET* that "the general answer" to the question "How many principles do we need?" is that "A principle *p* should be added to our list of principles if there exists at least one paradigm case for which *p* offers the best explanation of what one ought to do and why."[44] I also add the following remark: "The five geometrically construed principles articulated here are intended to be jointly sufficient for analyzing *all* cases related to new and existing technologies. This claim can, however, be understood in at least two

[37] I would like to thank Rob Reed for suggseting this helpful point to me.

[38] See, for instance, Gavagai.se.

[39] Shrader-Frechette (2017).

[40] Peterson (2017, pp. 37–38).

[41] Stewart et al. (1973, pp. 415–417), my italics.

[42] Kruskal and Wish (1978, pp. 30–31), my italics.

[43] Lokhorst (2018, p. 1).

[44] *ET*, p. 17.

different ways. First, it could be read as a stipulative definition. If so, the ethics of technology is, by definition, identical to the cases covered by the five principles. The second, and in my opinion more plausible interpretation, is to read this as a temporary conclusion that could be revised at a later point if need be. If we were to encounter new cases that could *not* be plausibly analyzed by the five principles, then it would be appropriate to extend the set of principles by a sixth or even a seventh principle."[45]

References

Anderson, M., & Anderson, S. L. (2014). GenEth: A general ethical dilemma analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.

Attfield, R. (2014). *Environmental ethics: An overview for the twenty-first century*. New York: Wiley.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Brown, C. (2011). Consequentialize this. *Ethics*, *121*(4), 749–771.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625).

Dafoe, A., & Russell, S. (2016). Yes, we are worried about the existential risk of artificial intelligence. *MIT Technology Review*.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge: MIT Press.

Goodall, N. J. (2016). Can you program ethics into a self-driving car? *IEEE Spectrum, 53*(6), 28–58.

Guardian Staff and Agencies, (2018). Tesla car that crashed and killed driver was running on Autopilot, firm says. The Guardian, March 31st, 2018.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). "The off-switch game", *arXiv preprint arXiv: 1611.08219*.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017a). "Ethically Aligned Design (EAD) - Version 2." Retrieved January 26, 2018, from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017b). "Classical Ethics in A/IS" Retrieved January 26, 2018, from https://standards.ieee.org/develop/indconn/ec/ead_classical_ethics_ais_v2.pdf.

Jonsen, A. R., & Toulmin, S. E. (1988). *The abuse of casuistry: A history of moral reasoning*. University of California Press.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. New York: Sage Publications.

Lokhorst, G. J. C. (2018). *Science and Engineering Ethics.*", 415–417. https://doi.org/10.1007/s11948-017-0014-0.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). "Should Robots be Obedient?". arXiv preprint arXiv.1705.09990.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289.

Paulo, N. (2015). Casuistry as common law morality. *Theoretical Medicine and Bioethics, 36*(6), 373–389.

Peterson, M. (2013). The dimensions of consequentialism: Ethics, equality and risk. Cambridge University Press.

Peterson, M. (2017). *The ethics of technology: A geometric analysis of five moral principles*. Oxford: Oxford University Press.

Peterson, M. (2018). The ethics of technology: Response to critics. *Science and Engineering Ethics*. https://doi.org/10.1007/s119.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532–547.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4*, 328–350.

Russell, S. (2016). Should we fear supersmart robots. *Scientific American, 314*(6), 58–59.

Shrader-Frechette, K. (2017). Review of the ethics of technology: A geometric analysis of five moral principles. Notre Dame Philosophical Reviews. University of Notre Dame. Retrieved November 11 2017 from. http://ndpr.nd.edu/news/the-ethics-of-technology-a-geometric-analysis-of-five-moral-principles/.

Stewart, A., Prandy, K., & Blackburn, R. M. (1973) Measuring the class structure. *Nature, 245*, 415.

Taylor, M. (2016). Self-driving Mercedes-Benzes will prioritize occupant safety over pedestrians, Retrieved January 26, 2018, from https://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians.

[45] Peterson (2017, p. 17).