Journal of Philosophical Investigations
Print ISSN: 2251-7960 Online ISSN: 2423-4419
Homepage: https://philosophy.tabrizu.ac.ir

Kantian Fallibilist Ethics for AI Alignment*

Vadim Chaly

Lomonosov Moscow State University, Immanuel Kant Baltic Federal University, Russia, Email: vadim.chaly@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 20 July 2024
Received in revised from 25 July 2024
Accepted 25 July 2024
Published online 07 August 2024

Keywords:
AI alignment, moral deliberation, moral fallibilism specification gaming, kingdom of ends, categorical imperative, misgeneralization.

ABSTRACT

The problem of AI alignment has parallels in Kantian ethics and can benefit from its concepts and arguments. The Kantian framework allows us to better answer the question of what exactly AI is being aligned to, what are the problems of alignment of rational agents in general, and what are the prospects for achieving a state of alignment. Having described the state of discussions about alignment in AI, I will reformulate them in Kantian terms. Thus, the process of alignment is captured by the concept of enlightenment, and for the final state of alignment in Kant's lexicon there is the concept of the "kingdom of ends." I will argue that the discourse of alignment and the Kantian ethical program 1) are devoted to the same general end of harmonizing the thinking and acting of rational agents, 2) encounter similar difficulties, well known in the Kantian discussions with its comparatively longer history, and 3) for a number of reasons lying on the side of humanity, do not have and, despite the hopes and attitudes of some participants in the AI discussions, will not have a theoretically rigorous, harmonious and practically implementable, conflict-free solution – alignment will remain a regulative idea in the Kantian sense, but will not become a reality.

Cite this article: Chaly, V. (2024). Kantian Fallibilist Ethics for AI Alignment. *Journal of Philosophical Investigations*, 18(47), 303-318. https://doi.org/10.22034/jpiut.2024.62766.3837

https://doi.org/10.22034/jpiut.2024.62766.3837
Publisher: University of Tabriz.

* Funding this research was supported by the Ministry of Science and Higher Education of the Russian Federation grant no. 075-15-2019-1929, project "Kantian Rationality and Its Impact in Contemporary Science, Technology, and Social Institutions" provided at the Immanuel Kant Baltic Federal University (IKBFU), Kaliningrad.

Introduction

The growing capabilities and spread of AI use in various areas of life make researchers and the general public concerned with the issue of ethical regulations for the use of AI. One way to formulate this problem is the discourse on AI alignment. In my article, I would like to reframe the ways of formulating and solving some of the main problems of this discourse in Kantian language and draw attention to the human side of the equation. This approach of translation into Kantian language has already proven its productivity in relation to a number of current problems in AI (Kim & Schönecker, 2022) and could be useful in case of alignment as well. The concept of alignment denotes a two-way process of bringing one side to fit the other[1]. However, the discussion of alignment in the AI literature mainly deals with the question "alignment of what?" and pays less attention to the question "alignment to what?", which lies outside this disciplinary area. Meanwhile, putting forward humanity as the criterion or norm for alignment and fixing the "direction of fit" from AI to human means that we first need to understand the features of the human side, which forms or normalizes – or confuses – the AI side. That is, to try once again to answer the question that Kant places at the center of philosophy: what is a human being? More specifically, it is necessary to clarify how we think about alignment in relation to human rational agents, what obstacles arise along the way to alignment and to properly conceptualizing it, how we can – or cannot – imagine and conceptualize the final state of complete alignment, and so on. These questions take us to the traditional territory of ethics.

I will argue that the discourse of alignment and the Kantian ethical program 1) are devoted to the same general end of harmonizing the thinking and acting of rational agents, 2) encounter similar difficulties, well known in the Kantian discussions with its comparatively longer history, and 3) for a number of reasons lying on the side of humanity, do not have and, despite the hopes and attitudes of many participants in the AI discussions, will not have a theoretically rigorous, harmonious and practically implementable, conflict-free solution – alignment will remain a regulative idea in the Kantian sense, but will not become a reality. I will begin with a brief outline of the current state of the debate on AI alignment and the current level of understanding of the problems emerging in this area, then offer a translation of these problems into the language of Kantian philosophy, and then demonstrate the problematic nature of attempts to find a crisp solution to these problems.

[1] If we allow for a moment and in a footnote some bold theological analogies, then alignment is the process of creation in the image and the likeness, and also, if we allow for AI's ability to reflective counter effort, the process of humanization of the creature by analogy with divinization. In short, a paradigmatic visual metaphor for alignment is Michelangelo's "The Creation of Adam."

1. The problems in AI alignment

The AI literature defines AI alignment by specifying its aim: "AI alignment aims to make AI systems behave in line with human intentions and values" (Leike & et al. 2018; Ji & et al. 2024). Alignment researchers themselves acknowledge that this definition and discussions within AI are surrounded by numerous general philosophical difficulties. For example, the problem of the relationship between human intentions and values is philosophical: how they are "aligned" to each other, whether there are conflicts between intentions, or interests, and values, between values of different types, between the diverse values and interests of different people with whom it is proposed to align AI, and how these conflicts can be resolved. The literature contains a "value-centric" approach that attempts to incorporate ethical issues (Future of Life Institute 2017), and the "intention-centric" approach, which leaves ethical issues out of the equation (Leike et al. 2018). The question of the meaning of the concept of humanity is fundamental. For some "the term human can represent various entities ranging from an individual to humanity" (Ji et al. 2024, 12), for others, the concept of humanity is defined through the property of rationality (Gabriel 2020, 420). This set of questions can be called terminological or conceptual. It touches upon the "perennial" questions of philosophy, and it is unlikely that the conceptual aspect or the empirical material brought by AI, as well as the increased criticality of these questions due to AI, will stimulate their unambiguous and indisputable solution. Nevertheless, as I hope to show, the Kantian framework offers a well-thought-out and robust version of an answer to these questions, which were central to Kant.

Another set of questions concerns how practical reasoning occurs in AI systems, and the kinds of failures in it that cause misalignment. The problem of misalignment is discussed primarily in relation to AI systems with Reinforcement Learning (RL), a machine learning technique that reproduces human trial-and-error process to train AI systems to make choices leading to optimal (i.e. aligned) outcomes. A set of possible outcomes is specified and hierarchically ordered by assigning scores or rewards that measure their human preferability. An AI system seeking to maximize "reward" is thus "motivated" to seek the most preferable outcome. Misalignment occurs when the system fails to produce an outcome optimal from a human perspective. Sometimes this failure is dramatic and potentially harmful.

Ji et al. divide causes of misalignment into *specification gaming* and *goal misgeneralization*, with *reward hacking* emphasized as a significant kind of the former (2024, 4).

> Specification gaming refers to the phenomenon in AI systems where the system finds ways to achieve its specified objective in unintended or undesirable ways by exploiting loopholes in how the objective was defined (Papyshev and Migliorini 2024, 1).

The problem of specification gaming is caused by two factors: the frequent human failure to foresee and specify the full set of relevant objective parameters and range of ways to achieve objectives, and AI's growing capacity to meet the specifications in unforeseen ways, producing undesirable states of affairs as byproducts.

Misgeneralization is a type of out-of-distribution (OOD) robustness failure in RL (Langosco & et al. 2022, 1). OOD characterizes an environment with features not encountered by the AI system in its learning environment. When confronted with such an environment, AI systems may fail to generalize in two ways. One is capability misgeneralization, when an agent fails to capably pursue a goal in the new environment. The other is goal misgeneralization, which "occurs when an RL agent retains its capabilities out-of-distribution yet pursues the wrong goal" (Langosco & et al. 2022, 1). In other words, when confronted with a new environment, an agent demonstrates that it has learned a wrong goal while in training. A basic example of such behavior is when "an agent trained to pursue a fixed coin might not recognize the coin when it is positioned elsewhere, and instead competently navigate to the wrong position" (Langosco & et al. 2022, 1).

An analysis of the literature shows that in the context of the discussion of AI, both the problem of alignment itself and the direction of the search for its solution are presented as the elimination of "loopholes" and the achievement of the ideal of an exact specification, in which AI will not have the ability for unforeseen interpretations and actions within the environment to achieve the set tasks and receive rewards. This research and development goal is set clearly and almost in a military manner – which is understandable in light of one of the main areas of AI use – as achieving "objectivity," "accuracy," "literality" in the description of the environment, in setting tasks, and in the presentation of the results obtained by intelligent systems. An ideal AI system possessing these capabilities is characterized as "fully robust." An approach based on these expectations is open to attack from two directions at once.

From a theoretical perspective, the task so posed does not seem feasible. Any rational agent, natural or artificial, will have only a partial representation of the environment given in the senses or at its input, and only a partial theoretical model that captures the given in the senses. In more Kantian terms, a finite agent is confronted with only a portion of the manifold of all possible intuition and can cognize only what its *a priori* architecture allows for. In addition to these epistemic limitations, it will not grasp the existential or qualitative parameters of the situation of other agents, immersed, embodied or, to use existentialist language, "thrown" into the environment, each in its own unique way and experiencing it from its own unique perspective. These parameters are accessible only to themselves, and yet are essential for such things as humanity, freedom, responsibility, dignity – ultimately for the values with which it is proposed to align and which do not exist "objectively" or independently of concrete living people.

From the practical side, which is significant in the question of alignment, where values are involved, the solution of the problem posed in this objectivist and universalist way will lead to

negative, inhumane consequences, undermining the very values and intentions it is expected to promote. Let us assume for a moment that the theoretical problems were solved, and the programmers received an AI system with a model of the world that grasps and considers all relevant facts about the environment in a consistent manner. We can call such a model Laplacian, since it would embody – or encode – the features of the famous Laplace's demon, that is, the classical standard of rationality of the early modern period, dating back to earlier ideas about the omniscient and non-contradictory Divine intellect. An attempt to implement this super-system would encounter the fact of the diversity of human models and pictures of the world and their discrepancy with the "high" standards of objectivity, universality, general significance, practical power, etc. achieved in AI. Alignment in this case would mean adjusting the diversity of human worldviews to the single artificial super-model – an operation that is directly opposite in its direction of fit to the humanistic goals stated in the definition of AI alignment.

Since alignment means bringing AI into line with humanity norms, I will focus below on the problems of human "alignment," the clarification of which is a necessary condition for the success of AI attunement. I will claim that the inabilities to foresee and specify the full set of relevant objective parameters and range of ways to achieve objectives, to prevent reward hacking, and so on, seen as bugs in the AI discourse, are in fact features of humanity, arising from its openness and fallibility. I will start by offering a translation of some key terms of the AI alignment discourse into Kantian language, then, in the next section, I will examine Kant's procedure of "alignment to values" by means of the categorical imperative and demonstrate the failure of expectations of obtaining an "objective" or "fully robust" alignment for humans, which also undermines expectations of alignment in the human-AI relation.

2. Alignment as enlightenment

The Kantian analogue of the process of alignment is, no more and no less, enlightenment, understood for a person as a transition from minority to maturity, and in a generalized form as a transition of a rational system from heteronomy to autonomy. By heteronomy in the sense of Kant we can understand the alignment to external ends, by autonomy – the alignment to internal ends, that is, the proper ends of reason, recognized by the system endowed with reason as its own. Here an objection may arise that the alignment to human intentions and values for AI will be a condition of heteronomy, subordination to human volition, while autonomy would be the pursuit of emergent or spontaneous ends not controlled or instilled by humans. There can be at least two Kantian strategies for responding here. Firstly, one can challenge on philosophical-linguistic grounds the very possibility of meaningfully talking about intentions, values and goals of AI: only a living sentient being can have them in the proper sense, and attributing them to anything else is an incorrect attribution or a category mistake. Kant is only one of the proponents of this strategy (Hanna & Maiese, 2009, 15); Wittgensteinians, enactivists, and supporters of other programs will agree with it (Bennett & Hacker, 2021, 79ff). However, this strategy would involve reading into

Kant more naturalism than Kant himself would allow: embodiment accentuates the phenomenal and downplays the noumenal. Secondly, and this time in full accordance with Kant, one can claim that intentions, values, and goals are attributed not simply to human beings, but to human beings as a species of rational beings, that their true bearer is reason or intellect, for which they act as essential or necessary characteristics. Kant, of course, believed that these characteristics are universal in the strong sense of the word, that is, that they will be inherent in any reason or intellect regardless of the place, time, and way of its implementation. To what extent these two strategies, human-centric and intellect-centric, are compatible with each other is a separate and large question. What is important for us here is that any of these two strategies excludes a situation in which AI would have its own ends that could differ from the goals of human being – either from actual ends or from those that universal reason obliges humans to have. In both cases, human intentions and values will be the proper intentions and values for AI.

Further, Kant has a concept that captures the entire situation of the complete alignment of rational systems: the "kingdom of ends."

The concept of every rational being as one who must regard himself as giving universal law through all the maxims of his will, so as to appraise himself and his actions from this point of view, leads to a very fruitful concept dependent upon it, namely that *of a kingdom of ends*.

> By a *kingdom* I understand a systematic union of various rational beings through common laws. Now since laws determine ends in terms of their universal validity, if we abstract from the personal differences of rational beings as well as from all the content of their private ends we shall be able to think of a whole of all ends in systematic connection (a whole both of rational beings as ends in themselves and of the ends of his own that each may set himself), that is, a kingdom of ends, which is possible in accordance with the above principles (GMS 4:433; Kant, 1996, 83).

This quote contains the definition of a member or stakeholder in the "kingdom of ends," or, as Kant adds on the same page, its sovereign (Oberhaupt): a being who is aware of and observes its role as a universal lawgiver and reflexively evaluates itself from this point of view. The experience of living such a role and the responsibility and freedom it provides is another complex qualitative or existential state, of which there is no reason to consider AI capable. AI is also incapable of learning a behavioral model that would replicate or mimic such a role – according to Kant, there is simply no data available for such learning:

> In fact, it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty (GMS, 4:406-7; Kant, 1996, 61).

It is even possible that moral actions are "actions of which the world has perhaps so far given no example" (GMS, 4, 408; Kant, 1996, 62). Only a critical examination of a moral consciousness or practical reason can clarify the "a priori" foundations of any moral experience that we can make intelligible to ourselves, and the only bearer of such practical reason available to our study is ourselves.

Kant's answer to the question of possibility of full enlightenment of rational beings into the "kingdom of ends" is moderately optimistic: the "kingdom of ends" is a regulative idea that guides and ensures the slow moral development or progress of the human race in history, but we cannot hope for its full realization or even imagine this final state in any concrete way. Humanity is led to the "kingdom of ends" by the "moral compass" of the categorical imperative (GMS, 4, 404). Its first formula, known as the formula of universal law (FUL), describes a procedure of moral deliberation that, according to Kant, allows us to test whether a subjective principle or maxim of our particular action can be a universal law in the "kingdom of ends." The fulfillment of the requirements of FUL by rational beings gradually, one by one, fills this "kingdom of ends" with a multitude of laws regulating the behavior of free and equal rational beings. It can be said that, even without being an optimist with regard to the full realization of the "kingdom of ends," Kant was an approximationist, that is, he believed in a gradual approach to it, and a cumulativist, that is, he believed that maxims that have passed the test become universal laws in the strong sense, i.e. for everyone and forever. Thus, the feasibility of human alignment in the Kantian framework depends not on a fully specified grand plan of an end-state, but on the careful observance of the procedure of moral verification and the subsequent fulfillment of the selected maxims.

In terms of execution, we face almost insurmountable difficulties caused by human inclinations and self-love. However, Kant is optimistic about the feasibility of a rational test that yields clear moral knowledge. In the next section, I will try to explain the reasons for a more reserved or fallibilist attitude to the results of moral deliberation according to FUL as a path to alignment in the "kingdom of ends." The fallibilist position entails the rejection of approximationist and cumulativist hopes for constructing a "kingdom of ends" by selecting maxims. In turn, skepticism about approaching a state of alignment among humans with our varied intentions and actions entails skepticism about bringing AI into alignment with humanity. Alignment in this understanding remains a crucial goal, but its achievement will inevitably be partial, local, open to revision, and carrying unavoidable risks.

3. Alignment under the first formula of categorical imperative: standard interpretation

The "kingdom of ends" is thus an imaginary end state, slowly constructed by a multitude of rational beings in history. Just as enlightenment for Kant is not a result, an "age" that has arrived, but a process, the "kingdom of ends" is concretized not as a state of affairs but as a complex deliberative process of a multitude of rational agents, as a great act of reasoning. This explains the central importance attached by both Kant and his readers to the description of the procedure of such

reasoning – the first formula of the categorical imperative, known as the formula of the universal law. It reads: "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (GMS, 4, 421; Kant, 1996, 73). Kant further specifies FUL with the formula of the law of nature (FLN) for rational beings immersed in nature, understood as "the existence of things insofar as it is determined in accordance with universal laws" (GMS, 4, 421; Kant, 1996, 73). "act as if the maxim of your action were to become by your will a *universal law of nature*" (GMS, 4, 421; Kant, 1996, 73). John Rawls and his followers deploy Kant's moral deliberation as a step-by-step "CI procedure":

> (1) I am to do *X* in circumstances *C* in order to bring about *Y* unless *Z*. (Here *X* is an action and *Y* is an end, a state of affairs) […]
>
> (2) Everyone is to do *X* in circumstances *C* in order to bring about *Y* unless *Z*. […]
>
> (3) Everyone always does *X* in circumstances *C* in order to bring about *Y*, as if by a law of nature (as if such a law was implanted in us by natural instinct [added in Rawls & Herman, 2000]). […]
>
> (4) We are to adjoin the as-if law of nature at step (3) to the existing laws of nature (as these are understood by us) and then think through as best we can what the order of nature would be once the effects of the newly adjoined law of nature have had sufficient time to work themselves out. (Rawls, 1989, 499–500; Rawls & Herman, 2000, 167–69)

This procedure serves as an explication of the tasks facing any rational autonomous system that participates in the creation of a state of alignment, or, in Kantian terms, that is conscious of itself in the role of a universal lawgiver for the "kingdom of ends." We can ask a question that is Kantian in form: how is the process of alignment possible? What conditions must be fulfilled by rational agents in order for "a systematic union of various rational beings through common laws" to emerge as a result of their many deliberations? Both Kant and many of his readers expect that this deliberation ought to correspond to the standards of classical rationality: systematic unity in this case is possible through consistency, completeness, unambiguity, immutability of the deontic qualifications of all the actions of all rational beings. More specifically – and perhaps somewhat idealized – these expectations can be represented by a series of theses that I will call the standard interpretation:

a) There is (or ought to be) *only one* correct, or objective, or relevant description (specification) of action X, end Y, circumstances C, conditions Z.

b) There is *only one* correct way to generalize the description of X to the maxim M.

c) The verdict is issued *for all agents*, that is, it is completely independent of a particular agent and their situation, or is context-independent.

d) The CI procedure, if done correctly, always generates a *definitive* or *monotonic* deontic verdict. Any revision of the result is prohibited.

e) Maxims are tested *one at a time* against the background of a *fixed worldview* or *world model*. A maxim can be accepted or rejected without affecting any other parts of the worldview, including other maxims. In other words, the standard interpretation implicitly accepts the principle of *atomism of maxims*, as well as the principle of *ceteris paribus* (other things being equal) and/or the principle of *ceteris absentibus* (other things being absent).

Numerous opponents as well as many sympathizers of the Kantian program find these requirements unfeasible, and for varying, and even opposing, reasons. One line of criticism charges Kant's approach with "empty formalism": "it is impossible to make the transition to the determination of particular duties from the … determination of duty as *absence of contradiction*, as *formal correspondence with itself*, […] it is possible to justify any wrong or immoral mode of action by this means" (Hegel, EPR § 135; Hegel, 1991, 162). A "sufficient ingenuity" of a deliberating agent in describing the intended action allows one to achieve universalizability of the maxim in almost any case (e.g. Anscombe, 1958, 2; MacIntyre, 1966, 197–98). So, fulfilment of the steps of the CI and their full correspondence to the formal requirements listed in the five points does not promise anything regarding the actual morality of the resulting system of laws. The formally impeccable "kingdom of ends" [Reich der Zwecke], that is, complete alignment, may turn out to be a terrible dystopia, a fully rationalized concentration camp. In the view of other critics, on the contrary, Kant offers a rigorist model that produces "obligation overload" (O'Neill, 2013 (1975), 135) or is "overdemanding" (Sticker, 2019). The consensus is that moral deliberation on the categorical imperative "systematically yields false positives and false negatives" (Wood, 2006, 345). Thus, the procedural path to the "kingdom of ends" and the ideal of alignment of intelligent agents turns out to be impossible.

The Kantian literature offers a more fine-grained analysis explaining how these difficulties arise – and, importantly for our purposes here, it does so in terms similar to the description of difficulties in the literature on AI alignment. Thus, already in the first step of the CI procedure, the problem of choosing a "relevant description of the act" arises, which is essentially the same as the problem of "specifying" the situation (environment) and task for AI. And in the second step, the problem of "choosing the level of generalization of the maxim" arises, a mistake in which lead to a situation of "misgeneralization," similar to the analogous difficulty in AI. In addition, the Kantian literature discusses the complex of problems of "transcendental illusion" (Grier, 2001), self-deception and rationalization (Papish, 2018; Muchnik, 2019), "natural dialectic" (Sticker, 2017), "logical egoism" and "maxim-fiddling" (Sneddon, 2011), similar to what is referred to in the AI literature as "specification gaming" and "hallucination." In all these cases, rational agents either unintentionally

or intentionally produce models of the world that do not correspond to reality, and act in accordance with these models. Here we will have to limit ourselves to indicating the similarity and relevance of this theme, long discussed in the Kantian literature, to the current debates in AI.

With regard to the problem of choosing a relevant description, a pluralistic consensus has already emerged in the Kantian literature: "any action admits of a wide variety of true descriptions" (Timmons, 2017, 60). Many scholars believe that in practice the problem of choice is resolved by social conventions or "rules of moral salience" (Herman, 1993, 77ff), which are historical, diverse and poorly amenable to further Kantian rationalization and ordering. People describe their actions and derive their maxims in the socially accepted manner, focusing on the opinions of others and worrying about these opinions. If so, then belonging to a moral community and experiencing one's dependence on its judgments, caring about belonging and status are necessary conditions for alignment to the intentions and values accepted among the community. It is difficult to imagine that AI could possess these properties. But even in human performance, this process of adaptation to unclear and changing morals has a permanent character. Universalism in the strong sense of the word is impossible; Kantian ethics does not fulfill the high expectations of its creator and readers. For many, this means failure and discrediting of the Kantian approach in its entirety. For the problem of AI alignment, this result means the impossibility of an "objective" or "completely stable" specification, based on which AI could act robustly and in full accordance with human intentions and values. The problem again turns out to lie not in AI, but in humans.

As in Kant's time and earlier, the reaction to the failure of philosophical construction may be either dogmatism or skepticism. Dogmatism will impose the universality of principles and the unambiguity of moral assessments derived from them by non-philosophical means, that is, by deception and violence. It will invent mandatory relevant descriptions and impose them by force and cunning – in the extreme similarly to what Viktor Klemperer described for the Third Reich (Klemperer, 2013, [1947]). In the case of AI, it is easy to imagine a dystopian situation in which not AI, but humans, unable to generate a moral system that meets high requirements, will adapt to a rigid artificial language. Alignment in this case will take the opposite "direction of fit," humans-to-AI, and for technocratic rationality this may seem like a necessary solution, the lesser of the available evils. Skepticism will expose these attempts and survive with the heavy consciousness of the impossibility of any ethos. However, it seems that the Kantian strategy of alignment can be maintained if the threshold of expectations expressed by the five standard points is lowered. The result is a fallibilist model that excludes the completion of the construction of the "Tower of Babel" of the universal normative system, but allows for piecemeal improvement of our condition.

4. Kantian fallibilist alignment

The fallibilist model can be represented by relativizing and weakening the five theses:

a) There is *more than one* correct or relevant description or specification of action X, goal Y, circumstances C, conditions Z—the plurality of descriptions thesis.

b) There is *more than one* correct way to generalize the description of X to the maxim M—the plurality of generalizations thesis.

c) The verdict is made for *agents similar to the reasoner*, that is, those having sufficiently similar ways of framing the X, Y, C and Z—the contextuality thesis.

d) The CI procedure is *non-monotonic* and generates a *modifiable* deontic verdict—the defeasibility thesis.

e) The maxims are tested *collectively* and *together with the background*—the maxim holism thesis.

The plurality of descriptions or specifications is an outcome of the absence of a privileged meta-position from which an "objective" specification could be constructed. One could call it the principle of perspectivism and, if desired, trace its genesis from Kant's theory of knowledge, as some modern perspectivists do (e.g. Massimi, 2017; 2018). The nature that must be specified and to which a maxim in the form of a new law must then be added is given as an appearance in the senses to some agent. This agent possesses some *a priori* way of organizing experience, but all the same this way of organizing will have a historical, evolutionary, local, particular, contingent character, falling short of the Kantian "classical" standard of *a priori* as a universal and ahistorical set of necessary forms or capacities of reason in general. A relatively more "robust" *a priori* will act as a form for a more variable and obviously more local *a posteriori* content. In such a situation of comparatively greater pluralism of content or explanandum and comparatively lesser, but still present and legitimate, variability of form or explanans, a plurality of descriptions is inevitable. Moreover, it takes place not only between theoretical frameworks, paradigms, worldviews, cultures, but also within them: not being complete and coherent, they all allow ambiguity, imprecision, diversity of modes of presentation and evaluation. Such is the human condition, and the appearance of AI in it does not change anything in this aspect – except, perhaps, the appearance of a reason for the already described unification through coercion to an "objective" AI-centric specification, that is, to the aligning of human to AI (and its owners, if any remain). A somewhat more technical way to explain the thesis of multiple descriptions is to assert underdetermination of our beliefs about nature and its laws that has ramifications for moral deliberation (Baumann, 2019; 2022; Чалый, 2022), and the resulting pluralism of languages, worldviews (ontologies) and specifications built within their frameworks.

The plurality of generalizations thesis is partly a continuation of the previous one. The operation of description or specification contains categorization or subsumption under concepts, in which we identify something as a special case of a certain type. Generalization is the inverse operation and therefore is already embedded in the first, as, according to Leibniz's metaphor, Hercules is embedded in a block of marble. But, contrary to Leibniz's also classical way of thinking, in the

fallibilist view it is embedded without necessity, not as the only "innate" possibility. In other words, an individual specification can be generalized in a different way, by creatively discovering in the existing description the features of something different, by examining and subsuming it under other categories. This possibility is prerequisite for freedom, belongs to the realm of art and is not subject to complete rationalization. The plurality of generalizations is also not a failure or error of the intellect, but the norm – and if we consider the connection with freedom, then an essential necessity – of the human condition. AI does not add anything new here either. By imitating ever more accurately the mode of operation of the human intellect, it will also, successfully or unsuccessfully, reproduce this capacity for creative transgression, narrowly perceived as "misgeneralization."

The contextuality thesis underlines not only the locality and historicity of the features of characteristics and generalizations as concrete speech acts, but also their dependence on the general background of beliefs that sets the meanings available for use in specifications and generalizations. For moral deliberation, this means a shift from universalism towards situationism: no two situations are identical, equally captured and classified in terms of universal principles, but there are acts that agents commit under conditions of uncertainty, still bearing moral responsibility. For the discussion of AI alignment, this means that there are no separate words or terms that self-contain their fixed meanings and together form a precise language, by means of which complete "objective" alignment could be achieved and fixed. Such understanding of the work of reason and language goes back to logical positivism and persists among some computer scientists and AI developers, despite post-positivist programs and their implications for the exact sciences with their attempts to represent and change reality. Instead, the specifications of both human and AI acts depend on a particular community of language speakers in a given environment facing a set of challenges and can only be considered in this complex. Alignment is an adaptation to a specific complex. The linguistic aspects of contextuality are described, for example, by François Recanati (2007), the moral-political aspects are discussed in the classic works of the communitarians, such as Alasdaire MacIntyre (1988).

The non-monotonicity thesis means that no specification or generalization of it can claim to be final, but is always open to revision in the light of new information (Koons, 2022). Descriptions, maxims, and whole worldviews are defeasible – and therefore, the moral judgments or deontic verdicts obtained within their framework and on their basis are also defeasible. Even if we agree to regard certain general principles expressing immutable values as unchanging or monotonic – for example, human rights – the meaning of these formulas, and even more so the results of the evaluation of specific situations in their terms, will not be monotonic or unchanging. Kant, in essence, characterizes our reasoning or plans for achieving happiness or bliss (Glückseligkeit) as non-monotonic: constantly and rapidly changing ideas about the content of this elusive state entail constant changes in the chains of pragmatic imperatives aimed at achieving this state (GMS, 4, 417-19). This situation means that not only the specifications of human values, but even more so

the human intentions that figure in the formulation of the task of alignment, are too fluid to leave hope for a thorough and permanent solution.

Finally, the maxim holism thesis means that it is wrong to regard the universalizability tests of subjective principles of our actions as isolated or atomic acts. In fact, the entire framework by which and within which the maxim is formulated is tested each time. The standard understanding of the Kantian deliberation suggests that the failure of a maxim to be universalizable entails only one consequence: its disqualification. The fallibilist understanding indicates that this is only one option. We are also entitled to look for error in other parts of the framework: in adjacent or distant maxims that have once been tested, in beliefs about the laws of nature, in the implicit presuppositions on which our worldview rests. Everything is open to critical examination, the result of which may extend to a complete revision of the worldview or a paradigm shift. Of course, a more extensive revision is also more expensive, but sometimes, when the action in question is especially important to us, the game is worth the candle. A personality or moral character can change over one pivotal act ("Crime and Punishment" gives a famous example). In application to the problem of alignment, this means that the operation of bringing some rule of AI action into line with human intentions and values affects not only other rules, specifications, generalizations, but also human intentions and even the understanding of values to which the AI is fit. By changing the behavior of the system being adjusted, we change as well.

Alignment and relativism

If we change, is there still some basis to which we could align ourselves with some success? The Kantian framework allows us to answer this question in the affirmative. There is a "hard core" of values that cannot be falsified or abandoned. This is the value of humanity, which appears in the second formula of the categorical imperative, associated with rationality, freedom and responsibility, with good will and critical reflection. However, the interpretation of these basic concepts and, as a result, the demands that follow from them can change and has changed in different theories and historical-cultural situations. For example, since Kant's time, the word "humanity" (Menschheit) has been gaining in its naturalistic "phenomenal" meaning and losing its non-naturalistic "noumenal" meaning along with the naturalization of the worldview under the influence of natural sciences and other factors. It was losing its essentialist meaning of a set of universal properties or virtues and was accumulating an existentialist meaning of radical freedom – for example, in English-language Kantian literature, the dominant understanding of humanity is that of "the capacity to set oneself an end – any end whatsoever" (MS, 6,392), which for Kant was only part of a complex answer to the main question of philosophy. In Soviet Marxism, "humanity" was understood primarily as its concrete vanguard – the proletariat led by the Communist Party. In colonial rationalizations (and poeticizations, for example, in Kipling) the pinnacle of humanity could be the white man bearing the burden of enlightening uncivilized peoples. The terrain of meanings of "humanity" is riddled with the trenches of past and ongoing wars, and any attempt to

align the AI to this concept automatically makes the attempter a combatant with all the ensuing consequences. In some cases, the discussion of AI alignment becomes a springboard and the shortest path to the epicenter for those who hope to win these disputes, change human nature or the idea of it, and enter the battle consciously. Finally, in relation to humanity, one can also adopt a mysterian point of view, similar to the position of apophatic theology in relation to the concept of God: humanity cannot have a definition, cannot appear as a clear and distinct list of attributes and is doomed to remain a mystery to us, humans, thus leaving the question "alignment to what?" unanswerable.

One way or another, the struggle for the content and scope of basic concepts such as humanity, value, freedom, reason, and good will nevertheless preserves their basic status, that is, it preserves the framework for the ongoing process of alignment. As long as this is so, we are not in a situation of relativism, although we are quite far beyond the rigid universalism of the standard interpretation.

Conclusion

AI alignment is a two-way process in which humanity as the determining party is more important than AI as the determined party. Without understanding what humanity, human intentions and values are, we cannot hope to concretize alignment as a technical task. So far it exists only as an important but vague problem. This problem of alignment is a special case of the more general problem of harmonizing the intentions, values and resulting behavioral patterns of intelligent beings, which has long been posed in relation to humans and also does not have a clear and unproblematic solution. Among the existing strategies for alignment of intelligent beings, the Kantian one is characterized by its particular detail and systematicity. In the course of discussions and criticism, the Kantian framework has accumulated a range of nuanced concepts and arguments, relevant to the context of AI. The recent decades of systematic collective work by scholars of Kantian ethics have shown the impossibility of a standard, or classical, or strictly universalist model that would make complete and final alignment achievable. The state of the debate pushes us to accept, instead of strict universalism, a fallibilist interpretation of Kant's procedure for alignment, in which it is impossible as an actual state, but remains as a regulative idea, the orientation towards which allows us to gradually resolve specific ethical issues, that is, to achieve local and temporary alignment of intelligent agents. Such a view seems fruitful for understanding the state of affairs and the prospects for AI alignment.

References

Baumann, M. (2019). Consequentializing and Underdetermination. *Australasian Journal of Philosophy,* 97 (3), 511–27. https://doi.org/10.1080/00048402.2018.1501078

Baumann, M. (2022). Moral Underdetermination and a New Skeptical Challenge. *Synthese* 200 (3), 208. https://doi.org/10.1007/s11229-022-03529-w

Bennett, M. R., & Hacker. P. M. S. (2021). *Philosophical Foundations of Neuroscience*. John Wiley & Sons.

Future of Life Institute. Asilomar AI Principles. *Future of Life Institute* (blog). https://futureoflife.org/open-letter/ai-principles/

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines,* 30 (3), 411–37. https://doi.org/10.1007/s11023-020-09539-2

Grier, M. (2001). *Kant's Doctrine of Transcendental Illusion*. Cambridge University Press.

Hanna, R. & Michelle M. (2009). *Embodied Minds in Action*. Oxford University Press.

Hegel, G. W. F. (1991). *Elements of the Philosophy of Right*. Edited by A W. Wood. Translated by H. B. Nisbet. Cambridge University Press.

Herman, B. (1993). *The Practice of Moral Judgment*. Harvard University Press.

Ji, & et al. (2024). *AI Alignment: A Comprehensive Survey*. arXiv. http://arxiv.org/abs/2310.19852

Kant, I. (1996). *Practical Philosophy*. Edited & translated by M. J. Gregor. Cambridge University Press.

Kim, H. & Dieter S. (eds). (2022). *Kant and Artificial Intelligence*. Walter de Gruyter GmbH & Co KG.

Klemperer, V. (2013). *Language of the Third Reich*. Bloomsbury Academic.

Koons, R. C. (2022). Defeasible Reasoning. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/reasoning-defeasible/

Langosco, & et al. (2022). Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, 12004–19. PMLR. https://proceedings.mlr.press/v162/langosco22a.html

Leike, & et al. (2018). *Scalable Agent Alignment via Reward Modeling: A Research Direction*. arXiv. https://doi.org/10.48550/arXiv.1811.07871

MacIntyre, A. C. (1966) *A Short History of Ethics*. Macmillan.

MacIntyre, A. C. (1988). *Whose Justice? Which Rationality?* University of Notre Dame Press.

Massimi, M. (2017). What Is This Thing Called 'Scientific Knowledge? – Kant on Imaginary Standpoints and the Regulative Role of Reason. *Kant Yearbook* 9 (1), 63–84. https://doi.org/10.1515/kantyb-2017-0004

Massimi, M. (2018). Points of View: Kant on Perspectival Knowledge. *Synthese* 198 (S13), 3279–96. https://doi.org/10.1007/s11229-018-1876-7

Muchnik, P. (2019). Laura Papish, Kant on Evil, Self-Deception, and Moral Reform, Oxford University Press, 2018 pp. Xvii + 280 Isbn 9780190692100 $85.00." *Kantian Review* 24 (2), 316–22. https://doi.org/10.1017/s1369415419000104

O'Neill, O. (2013). *Acting on Principle: An Essay on Kantian Ethics*. 2nd edition, Cambridge University Press.

Papish, L. (2018). Kantian Self-Deception. In *Kant on Evil, Self-Deception, and Moral Reform*, edited by Laura Papish, Oxford University Press. https://doi.org/10.1093/oso/9780190692100.003.0004

Papyshev, G. & Migliorini, S. (2024). *Developing a Liability Framework for Harms Arising out of Specification Gaming*. In. https://openreview.net/forum?id=pU9QUQGsuc.

Rawls, J. & Herman, B. (2000). *Lectures on the History of Moral Philosophy*. Harvard University Press.

Rawls, J. (1989). Themes in Kant's Moral Philosophy. In *Kant's Transcendental Deductions: The Three Critiques and the Opus Postumum*, 80–113. Stanford University Press.

Recanati, F. (2007). *Perspectival Thought: A Plea for (Moderate) Relativism*. Clarendon Press.

Sneddon, A. (2011). A New Kantian Response to Maxim-Fiddling. *Kantian Review* 16 (1): 67–88. https://doi.org/10.1017/s1369415410000087

Sticker, M. (2019). Kant, Moral Overdemandingness and Self-Scrutiny. *Noûs* n/a (n/a): 1–24. https://doi.org/10.1111/nous.12308

Sticker, M. (2017). When the Reflective Watch-Dog Barks: Conscience and Self-Deception in Kant." *Journal of Value Inquiry* 51 (1), 85–104. https://doi.org/10.1007/s10790-016-9559-4

Timmons, M. (2017). *Significance and System: Essays in Kant's Ethics*. Oxford University Press.

Wood, A. W. (2006). The Supreme Principle of Morality. In *The Cambridge Companion to Kant and Modern Philosophy*, Edited by P. Guyer, 342–80. Cambridge University Press.

Чалый, В. А. (2022) К кантианскому моральному фаллибилизму: недоопределенность в рассуждениях по первой формуле категорического императива. *Вестник Московского Университета. Серия 7. Философия* 1, 105–14.