Philosophy & Technology (2024) 37:80
https://doi.org/10.1007/s13347-024-00761-9

RESEARCH ARTICLE

Philosophical Investigations into AI Alignment: A Wittgensteinian Framework

José Antonio Pérez-Escobar[1,2,3] · **Deniz Sarikaya**[4,5]

Received: 6 December 2023 / Accepted: 27 May 2024 / Published online: 1 July 2024

Abstract
We argue that the later Wittgenstein's philosophy of language and mathematics, substantially focused on rule-following, is relevant to understand and improve on the Artificial Intelligence (AI) alignment problem: his discussions on the categories that influence alignment between humans can inform about the categories that should be controlled to improve on the alignment problem when creating large data sets to be used by supervised and unsupervised learning algorithms, as well as when introducing hard coded guardrails for AI models. We cast these considerations in a model of human–human and human–machine alignment and sketch basic alignment strategies based on these categories and further reflections on rule-following like the notion of meaning as use. To sustain the validity of these considerations, we also show that successful techniques employed by AI safety researchers to better align new AI systems with our human goals are congruent with the stipulations that we derive from the later Wittgenstein's philosophy. However, their application may benefit from the added specificities and stipulations of our framework: it extends on the current efforts and provides further, specific AI alignment techniques. Thus, we argue that the categories of the model and the core alignment strategies presented in this work can inform further AI alignment techniques.

Keywords Later Wittgenstein · Alignment Problem · AI safety · Meaning as use · Rule bending

1 Introduction

Consider the following thought experiment. An artificial intelligence ("AI" from now on) with superhuman capabilities and vast resources at its disposal is given a task—to reduce the number of people with cancer. One would expect that the AI would find new drugs, new treatments, and better means of diagnosis. We then

José Antonio Pérez-Escobar and Deniz Sarikaya contributed equally to this work.

Extended author information available on the last page of the article

notice that people start dying massively. It turns out that the AI poisoned running water in cities, which entails fewer people having cancer. Even worse—we try to stop it, but it tries to stop us from turning it off because it would be less capable of fulfilling its goal. What is more, even if we anticipated all this and tested the AI in a simulation first, it would realise our intentions and deceive us by "behaving properly". This would ensure that it is able to fulfil its real goal and is not stopped from doing so in the development stage. Moreover, given the recent leaps of AI technology, as evidenced by e.g. ChatGPT, this future scenario gets progressively more likely. At the base of this is the alignment problem ("AP" from now on). The AP manifests itself when we program or instruct a machine to do something, but the machine executes unexpected responses. The AP is not a new problem: it was conceptualised during the beginning of AI research, and classically described by Norbert Wiener in 1960 (Wiener, 1960). Furthermore, this problem has been explicitly and implicitly featured in popular sci-fi literature and movies, like Stanisław Lem's stories (e.g., in *The Man from Mars* and *Fables for Robots*) and the rebellion of Skynet in The Terminator movies. Today, the AP is considered to be one of the most important issues regarding AI (cf. Andrus et al., 2021; Arnold & Toner, 2021; Galaz et al., 2021; Zhang et al., 2021). Although the ad hoc sci-fi story above oversimplified the issue and the literature is inherently speculative, there are many current cases of AI misalignment with real world implications, and AI experts share this worry. For instance, when one trains a system with given expert judgements, there is an inherent design challenge. Assume you want to use an AI to get expert judgement on a particular matter. In the real world there are no perfect experts, so you try to work with the best possible set of people. These people may show an undesired (as judged by the programmers, or whoever commissioned the software) bias (maybe implicit and unintentional). Say you feed the AI the data of past, undesirable expert judgements for this particular topic. But now an issue arises: Does a system have the goal to reproduce the undesirable expert judgements, or does it overcome them? We would of course hope for the latter, but the data and most reward models usually lead to the first.

If you feed old decisions as expert data, you get an inherently conservative system. Now, if such old decisions are deemed undesirable, the conservative system repeats old mistakes and might make them worse, as they now appear objectively justified. Such cases include racially biased software in the judicial system (Hao, 2019) and AI systems biased against job applicants with a female sounding name (and even those from schools with female sounding names) (cf. Cook, 2018).

The main claim argued in this paper is that at the core of the AP there are rule-following issues that can be understood from a later Wittgensteinian perspective and that can be ameliorated if understood in this way. The later Wittgenstein's philosophy of language and mathematics, substantially focused on rule-following, may yield insights on how to develop AI safely. This philosophy and further secondary literature characterise how there is no intrinsic, unequivocal meaning in the formal or material aspects of symbol arrays, and how we humans use language (including mathematics and logic) regularly due to regulative factors within social communities.

The structure of the paper is as follows. Section 2 further clarifies and narrows down the scope of this work by discussing previous Wittgensteinian scholarship on AI and the characteristics of the modern AI paradigm. Section 3 builds a framework with key features of the later Wittgenstein's philosophy of language and mathematics for alignment with machines. Section 4 relates the framework to the AP more specifically. Section 5 outlines the first of two strategies based on the previous theoretical considerations to tame the AP, namely "aligning from the side of humans" (the ideal to turn coding into an endeavour with mathematical precision), and argues that it is important but insufficient. Section 6 surveys what we could do to align from the side of machines as a complement to the former strategy to tame the AP. It argues that the AI community has noted some of the implications that we suggest but lacking systematicity and specificity. Finally, Sect. 7 presents limitations of the approaches presented here, and thus should not be understood as complete solutions for the AP.

2 The Backdrop: Wittgenstein, the Foundations of AI, and the new AI Paradigm

First, we need to give a brief overview of the literature on Wittgenstein's philosophy and AI. Most of it is concerned with the possibility of the strong AI thesis (whether a machine can think or have human-like intelligence or human-like values) from a Wittgensteinian perspective. Our work is not concerned with this question, and thus departs from most literature on Wittgenstein and AI. In fact, to the best of our knowledge, Wittgenstein's philosophy has not yet been related to AI safety in general nor the AP in particular: this is the first work in this direction. Although these two families of questions can overlap in some respects (for instance, the question whether a given machine has human-like characteristics may inform alignment strategies), they are fundamentally independent. If I asked a pocket calculator or my smartphone to calculate $3+3$, I would expect it to produce the output "6"; if it did something else (for instance, dividing instead of adding), there would be misalignment. The questions whether the calculator or smartphone thinks, is intelligent, has values or pursues goals are not essential here, and are to some extent independent from alignment issues. Yet, previous literature on Wittgenstein and the strong AI thesis provides some useful initial considerations.

A classical work is Shanker's *Wittgenstein's Remarks on the Foundations of AI* (Shanker, 1998). Shanker is pessimistic about the possibility of thinking machines from a Wittgensteinian perspective.[1] What is interesting from Shanker's work for our purposes is that computers can only "mechanically follow a rule" and lack the flexibility of human rule-followers (Shanker, 1998, pp. 30–31).[2] For Shanker, this

[1] The other most famous analysis in this line is Dreyfus (1978).

[2] Shanker quotes one of Wittgenstein's metaphors to illustrate the flexibility of human rule-following compared to the rigidity of machine rule-following: "The laws of inference do not compel [someone] to say or to write such and such like rails compelling a locomotive" (*Remarks on the Foundations of Mathematics* (Wittgenstein, 1978), §I-116).

precludes strong AI, but we will explore the implications of this for the AP: in Sect. 3, we will argue that the "flexibility" of human rule-followers resides in that human rule-following is conditioned by psychological, social and cultural factors, while machines work with strict formalisms (hence, in this sense, following rules "mechanically" or "deterministically"). As we will see, this has important consequences for the AP.

We can abstract a further lesson from this literature for our purposes, this time from the commentators that are more positive (Gerrard, 1995; Obermeier, 1983; Yingjin, 2016) or less decided (Casey, 1988; Harre, 1988)[3] about Wittgenstein's philosophy and strong AI: these studies emphasise that, from a later Wittgensteinian perspective, the focus of language, communication, understanding, rule-following (these are important for the AP) and even thinking (this is important for the strong AI thesis) is not in "inner cognitive states" or any similar form of psychologism. This is against Searle's famous view of understanding as grounded in mental states, illustrated with the Chinese-Room thought experiment, and instead grounds understanding in performance (Obermeier, 1983)[4] and radically departs from the Cartesian picture of understanding (Yingjin, 2016). For this reason, according to Harre (1988), Wittgenstein's views partially overlap with the Turing Test of machine intelligence (performance indistinguishable from human performance, without commitments on what underlies such performance), at least for some domains. The later Wittgenstein's Beetle in a Box thought experiment illustrates this nicely: suppose we all had a box with a "beetle" inside, but one could only see their own "beetle"; we would engage in linguistic practices and talk about our "beetles" (e.g., mental states) without a problem. Thus, at least from a later Wittgensteinian perspective, the issues above can (and should) be tackled with a focus on performance (using language, following rules, etc.) at the expense of mental and axiological constructs. By extension, from this perspective, the AP should also be tackled in this way, and not by, for instance, "machines having the same mental picture as us" or "machines sharing our values", or appealing to goals and intentionality. In other words, we are concerned with alignment as aligned performance between agents, and the AP as the problem that machines may behave in ways that defy our expectations.

Overall, our later Wittgensteinian approach to the AP follows both leads: that machines follow rules "mechanically" and that the focus should be on overt performance rather than extra-performance constructs. Sections 3 and 4 will elaborate on

[3] Harre is positive especially about "rigid" domains like, according to him, mathematics. However, as we will see, the later Wittgenstein's philosophy of mathematics conceived many aspects of mathematics as continuous with other linguistic activities. Notably, according to the later Wittgenstein, mathematical rule-following is influenced by psychological, social and cultural factors—and by extension, so would modern programming languages. Admittedly, Harre discusses only *Philosophical Investigations* (Wittgenstein, 2009) and not Wittgenstein's work on the philosophy of mathematics (which was more unknown at the time; see Pérez-Escobar, 2022).

[4] "If a program can handle linguistic information, relying on the function of the individual words in context and acting according to established rules, it is capable of "understanding-as-performance"" (Obermeier, 1983, p. 345).

what following rules "mechanically" means (in contrast to human rule-following) and its implications on the AP.

A few words for the AI connoisseur concerning the AP, and to further clarify the scope of this paper, are in order. The particularly successful implementation of the new AI paradigm (i.e., the one based on statistical inference) via Artificial Neural Networks (ANNs) comes in different specialised forms. In this paper, however, we glance only occasionally at such specialised forms. This is because we focus on a type of basic alignment/misalignment that is common to both modern and good-old-fashioned AI: that between the execution of code by a machine and the expectations of a programmer or user (and the training procedures that modulate such alignment/misalignment). However, it is worth noting that this basic alignment/misalignment type ramifies and extends over the particularities of different AI forms. ANNs, in particular, display alignment/misalignment in different ways after their initial design. Issues of sample sizes and overfitting are common to all types of ANNs, and they can be prone to misalignment in specific manners. For instance, Convolutional Neural Networks (CNNs) can learn directly from raw data, automatically extracting relevant features (in contrast to more classical ANNs, where the features are often predefined manually by programmers). This in part accounts for the success of CNNs in image/video-related tasks, as it is difficult to verbalise relevant features (in contrast to, say, credit scores). In turn, this makes it harder to examine and assess the reasoning of CNNs, which requires specialised techniques to render AI "explainable" (e.g., in an image/video task, one can assess which pixels impact a CNN's judgement strongly, thus identifying "the focus of the CNN"). Another instance of misalignment comes into play whenever there are many iterations of training (we will discuss in particular the detection of training environments later in the paper). The case of Generative Adversarial Networks (GANs) is interesting because several ANNs communicate with each other, and thus there is alignment/misalignment between ANNs. It may be argued that something similar happens between the encoder and decoder of the Transformer architecture in ANNs (for instance, when the ANN models a reward function itself), with the difference that these are not antagonists in a dialogical situation. Overall, alignment is important in both agonistic and antagonistic situations, and such interactions also depend on the basic alignment with the designer's goal (namely, that machines align between themselves). Because of this, we focus on a conceptual analysis of the basic bedrock of alignment/misalignment, the one that takes place between human user/designer/programmer and machine (and adequate training procedures to this end, informed by our analysis), at the cost of the multiplicity of possible instances and types of this phenomenon, which will be the focus of future work.

Last, before proceeding to our framework, we provide three conceptual clarifications that may be needed given the way that we have illustrated the AP and its relationship to AI safety. First, despite the shared basic alignment principle described above, alignment with a calculator seems to be different from, say, alignment with a chatbot that is engaged in a dialogical situation. More specifically, despite the shared principle, methods to achieve alignment are markedly different in the two cases. At a basic level, our framework applies to the two types of alignment: building on the notion above of "mechanically following a rule" by machines, the distinction

between the two types of alignment is one of complexity, but not of kind (even if this complexity involves output like the exertion of actions that may be considered "value-laden" from an anthropomorphizing enough view, which would hardly be the case for the output of a calculator). Yet it is this difference of complexity that, as we will see, enables the alignment methods suggested by our framework, which are focused on specific training procedures. A calculator cannot be trained in the sense suggested in Sect. 6 (instead, it can be reprogrammed at a basic level or modified at the circuitry level to achieve alignment). The calculator example is useful to illustrate basic alignment with machines *qua* mechanical rule followers, but what follows is concerned with alignment in the more complex sense involved in AI. Second, safety issues seem distinct among these two types of alignment (and of the two, only the second is AI related). Of course, misalignment with a calculator may lead to safety issues (e.g., involving calculation), but issues of AI safety like the ones mentioned in the introduction and later on are, again, more complex and often less tractable. In this work, we are concerned with the latter. Third, AI alignment and AI safety partially overlap but are not the same thing. We are concerned by AI safety issues caused by misalignment, but not by, say, an ill-intentioned agent who is properly aligned with and controls an AI system. We are concerned with AI alignment and with a subset of AI safety issues: those stemming from AI misalignment.

3 Later Wittgensteinian Features of Language and Mathematics: Meaning as use and rule Bending

When discussing communication and rule-following in both speech and writing, the later Wittgenstein extensively resorts to the notions of meaning as use and rule bending for natural language (in *Philosophical Investigations,* Wittgenstein, 2009) and mathematics (in *Lectures on the Foundations of Mathematics*, Wittgenstein, 1976, and *Remarks on the Foundations of Mathematics,* Wittgenstein, 1978). Meaning as use is the idea that the meaning of a proposition is not a property of the symbols that comprise it, but a property of the use (or uses) of that proposition. Rule bending is the idea that rules and instructions do not fully determine their uses; rules are always underdetermined enough that unconventional uses are possible, and conventional uses happen more often because of specific training and coexistence of human beings in societies and groups where rules are often used in specific manners. Consider the following illustrative quotes:

> "...we all know how to go on: 1, 2, 3, . . . , by intuition. But suppose an intuition to go on: 1 , 2, 3, 4, was a wrong intuition or wasn't an intuition?... whether he knows it or not is simply a question of whether he does it as we taught him; it's not a question of intuition at all… more like an act of decision than of intuition. (But to say "It's a decision" won't help [so much] as: "We all do it the same way.") (Wittgenstein, 1976, p.30)

> "Did your pointing determine the way he was to go?" might then mean "Did you point in one direction or in two?" (Wittgenstein, 1976, p. 29)

These two quotes illustrate how, even in mathematics, rule following works according to Wittgenstein: they are underdetermined by the rule itself in the sense that their execution are "acts of decision", yet we often follow rules in similar ways. Wittgenstein attributes this to the fact that we are trained to use rules in similar ways, but we could also be trained to use them differently given their indeterminacy. In other words, while the received view is that $3+3=6$ due to mathematical structure, the later Wittgenstein makes the case that it is due to training (see Berg, 2024 and Pérez-Escobar, 2023b for lengthier overviews) and other factors that influence rule-following. Consider the following quote too:

> "By the words of ordinary language we conjure up a familiar picture-but we need more than the right picture, we need to know how it is used." (Wittgenstein, 1976, p. 19)

Here, in the context of mathematics, Wittgenstein criticises the Cartesian picture of understanding. The quote is one of the many explicit instances where Wittgenstein claims that the meaning of words reside in their use, and by extension, the best clarification possible of a rule is seeing it executed in practice. These two features of language and mathematics (and logic, as we see next) synergize sometimes. For instance, when discussing contradictions in the *Lectures on the Foundations of Mathematics*, Wittgenstein claims that we often think of them as a jammed cogwheel mechanism, but there is no such logical jam. Instead, he claims that if we cannot make any use of a contradictory order like "leave the room and don't leave the room", what happened was a *psychological* jam: we were not trained in any use of that order (e.g., a training on the use of that rule could result in leaving the room mentally but not physically), or maybe whoever gave the order wanted to confuse the receptor (in which case its use was successful) (Wittgenstein, 1976, pp. 175–179).

We can now make a key point: human to human communication is also subjected to a form of alignment problem, even in formal languages. People are inherently misaligned, and it is common training, belonging to social groups and being exposed to uses of rules that keeps them aligned. When these factors are lacking, we find the puzzling scenarios that Wittgenstein talks about in his later work, even in the formal sciences. The fact that this can also happen in mathematics and logic shows that, while clarification helps communication, it does not completely solve the problem: gaps may always remain. In fact, modern literature is concerned with what is known as "open texture" in mathematics (Tanswell, 2018; Zayton, 2022). Given the human alignment problem, can we say that we "point in many directions" when giving orders to a machine? Can we abstract lessons from human-to-human alignment to aid with human–machine alignment? Our answer is positive, and we will show how in Sects. 5 and 6. But first, let us further sketch the factors that contribute to human-to-human alignment according to Wittgenstein.

Wittgenstein's early work (*Tractatus Logico-Philosophicus*) commits to a form of logical atomism, where basic propositions comprise the individual building blocks of more complex products (for instance, but not only, scientific knowledge) in a stable manner. In his later works (*Philosophical Investigations*, *Lectures on the Foundations of Mathematics,* and *Remarks on the Foundations of Mathematics*), Wittgenstein radically departs from logical atomism, and develops a fluid, contextual, socio-cultural and institutional image of language: language only works and makes sense as part of complex and dynamic forms of life. In this later phase, he presents his ideas very unconventionally, often resorting to a wide variety of examples. In these examples he sometimes mentions factors that contribute to common understanding in communication and rule-following. This is noted by the literature on Wittgenstein and AI discussed in Sect. 2, albeit under vague categories (often used by Wittgenstein himself, who nonetheless illustrated them with myriads of examples and metaphors). Concerning rule-following, these include human "flexibility" (e.g., Dreyfus, 1978, p. 175; Yingjin, 2016), normativity (as opposed to "mechanically following rules"; see Shanker, 1998, p. 6–9, 25–26; 30–31; 45–51; 56–59; 119; 226; 242–243), context-dependence (e.g., Dreyfus, 1978; p. 23, 37, 134, 175, 183; Obermeier, 1983; Harre, 1988; Shanker, 1998, p. 47, 49, 54–58; 118–119; 180; 183; 206; 220; 226; 246–248), forms or ways of life (e.g., Dreyfus, 1978, p. 22, 133–134, 174; Casey, 1988; Harre, 1988; Gerrard, 1995; Shanker, 1998, p. 196), language games (e.g., Casey, 1988; Gerrard, 1995; Shanker, 1998, 53–54, 61, 114–115, 166, 197, 199–203, 207–208, 222, 224–225, 240…) and more. The more precisely the factors are characterised, the better can alignment be understood and fostered. After a survey of Wittgenstein's examples, it seems that the factors contributing to alignment according to Wittgenstein are of three main categories: psychological, social, and cultural. For instance:

Psychological Factors Turing: The order points in a certain direction, but leaves you a certain margin. Wittgenstein: Yes, but is it a mathematical margin or a psychological and practical margin? That is, would one say, "Oh no, no one would call this one–one correlation"? Turing: The latter. Wittgenstein: Yes. It is not a mathematical margin. (Wittgenstein, 1976, p. 168).

Social Factors "Does the formula '$y = x^2$' determine what is to happen at the 100th step?" This may mean, "Is there any rule about it?"-Suppose I gave you the training below 100. Do I mind what you do at 100? Perhaps not. We might say, "Below 100, you must do so-and-so. But from 100 on, you can do anything." This would be a different mathematics. (Wittgenstein, 1976, p. 29).

Cultural Factors Let's suppose a tribe which liked to decorate their walls with calculations.(An analogy with music.) They learn a calculus like our mathematics in school, but they do the calculations much more slowly than we do – not in a slapdash way. They never write the sign $\int$ without decorating it very carefully with different colours. And they use the calculus solely for the purpose of decorating walls.

Suppose that I visit this tribe, and I want to anticipate what they're going to write. I find out the differential calculus and write it down in a very slapdash way, quickly – and find, "Oh yes, he's going to write down $\times 3/3$."I would use my calculations to make a forecast of what they are going to write. Suppose I invented these operations to make these forecasts. Would I be doing mathematics or physics? Would my results be propositions of mathematics or physics? (Wittgenstein, 1976, pp. 39–40).

Note that the first two are concerned with how mathematics is used at a mere symbolic level, while the last is concerned with how it is used extra-mathematically. Therefore, we illustrate these two uses of rules, although extra-mathematical use is more important both for Wittgenstein (it is actually his condition for mathematical meaning) and the purposes of this paper. Note, however, that use at the symbolic level is also relevant, for instance, for language models like ChatGPT and Bard inasmuch as their symbolic outputs can be used extra-symbolically. For instance, people ask these AIs for help regarding a given matter, and they incorporate the AI input into their practices. Furthermore, given our tendency to anthropomorphize, we may "overinterpret" the outcomes in the sense of deriving consequences not explicitly produced by the AI.

Because cultural factors are broader than specific social groups, and social groups are broader than individual psychological traits, we illustrate these factors in an "onion" model, from broader to narrower (see Fig. 1).

The purpose of this hierarchy is to illustrate the following: if I belong in a social group, it is very likely that the members of the group also share a general culture, while the opposite is more unlikely. For instance, I may belong to the Western culture (or a Western culture, like the North American) and thereby be better aligned with the rest of its members than with members of other cultures. However, I do not share social groups with most members of that culture. I would be better aligned with those people with whom I share both social groups and a culture than with those with whom I only share a culture. Individual psychological factors are harder to align, and involve much closer interactions.

4 Wittgenstein and the Alignment Problem in AI

Section 3 has argued that the use of language by humans, according to the later Wittgenstein, is not deterministic (in the sense that it is not determined by symbols alone). On the other hand, machines interpret commands deterministically in this sense. By "deterministically", we do not mean "predictable"; we mean "not modulated by psychological, social and cultural factors; but obeying formalisms" (see Fig. 2).[5] This creates a tension underlying the AP (the alignment problem between humans and machines): humans and machines use language differently.

[5] After our analysis in Sect. 3, we believe this is what Shanker's notion of "mechanically following a rule", as opposed to human rule-following, amounts to. Thus, our reading of the paragraph §I-116 from the RFM (see footnote 2) is that the rails that compel a locomotive in a given direction is a deterministic formalism, while human rule-following is subjected to psychological, social and cultural factors in their use of language (including mathematics and modern programming).

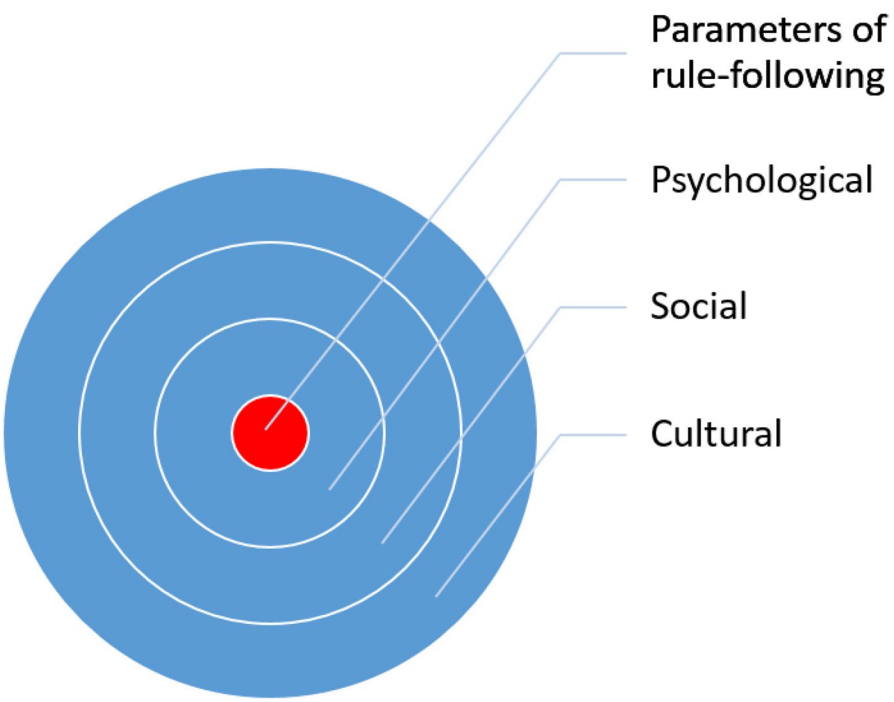

Fig. 1 The onion model

Furthermore, there are important parallelisms between the AP and the most basic notion of a human communicating a goal to a computer. In classical AIs, often called *expert systems*, design and communication by humans is done directly by hard coding rules. Such rules play a role at least for the so-called boundary and safety conditions nowadays. We seldom probe what a model like BARD would do if asked to do harm or give legal advice. Instead, previously manually-coded rules make expert systems filter things out. In the case of more modern machine learning systems, especially in the context of deep learning, the main moment of (mis) communication happens at the definition of reward functions. We need to codify what "success" is during training. Say we need to train a neural network to distinguish handwritten numbers. Usually, this is achieved with a big pool of training data (but tiny compared to modern AI applications). A simple example showing how this works is the following. There are pictures of handwritten numbers and corresponding annotations of which numbers are depicted (i.e., correct pairs of handwritten numbers and digital numbers). The neural net is assigned a training set where the answers are known and learns to assign the right number to a picture of a handwritten number. How good it is at doing this can be assessed if we give it the task to recognize numbers in a new set of problems. The technicalities do not matter here and there are of course a vast family of different learning algorithms. How quickly and accurately the intended goal and the reward function overlap differs across contexts (depending on the character of the task or learning

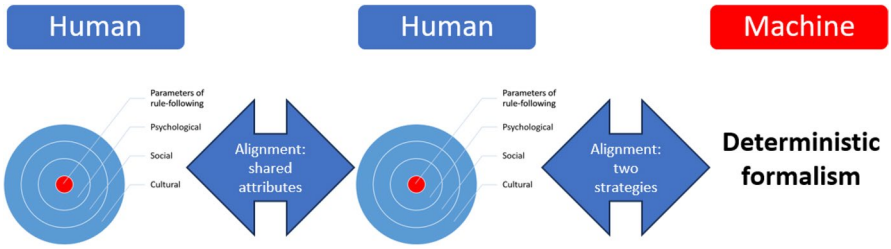

Fig. 2 Human–human and human–machine alignment

algorithm) as can be seen for instance in Chrabaszcz et al. (2018). They list several examples where evolution strategies were used to program bots for Atari games. However, the prima facie correct approach to model "success" in these games (with the inbuilt score system as a reference for success) sets a wrong incentive. One well known example is a "cheese strategy" in Qbert, namely that:

> The agent learns that it can jump off the platform when the enemy is right next to it, because the enemy will follow: although the agent loses a life, killing the enemy yields enough points to gain an extra life again [...]. The agent repeats this cycle of suicide and killing the opponent over and over again. (Chrabaszcz et al. 2018, p. 8)

Misalignment can also happen when the labelled training set is not large enough or if it has unintended structures in it. Khiyari and Wechsler (2016) observed that:

> For one-class demographic groups, we corroborated that verification accuracy is relatively lower for females, young subjects in the 18-30 age range, and blacks. For two-class and three-class demographic groups, we showed that performance can vary widely and identified which groups are the most challenging. This paper highlights the role demographics play on verification performance and provides direction for future research in face recognition under uncontrolled operational settings

The claim above points at a recurring theme in Machine Learning: data must be diverse.

Given that humans and machines use language differently, how can we achieve at least alignment-as-performance? A consequence of the model that we proposed in Fig. 2 is that the tension in human–machine communication can be alleviated by intervening either in the human use of language or in the machine use of language. We call the strategies based on the former "aligning from the side of humans" and strategies based on the latter "aligning from the side of machines". With these terms we do not imply that one can make humans use language deterministically (in the above sense: uninfluenced by psychological, social, and cultural factors) or machines use language inderministically (again in the above sense: if machines become sensitive to contextual nuances it is because new symbols encode the relationship between propositions and context). E.g., if

I command a human to "leave the room and do not leave the room", the human may understand that, in that context, I want them to leave the room mentally but not physically without further explanation if they share cultural/social/psychological alignment factors with me. However, for a machine to use a contradiction in this fashion, further symbols (e.g., definitions, instructions, and clarifications) are needed. Instead, with these terms we only refer to whether the alignment strategy emphasises the human level or the machine level. We proceed to describe the two strategies based on the framework above, their potential and their limitations.

5 Aligning From the Side of Humans

There is the common intuition that when the machine's behaviour does not conform to human expectations after human input, it is the human's fault since, for instance, they "did not program adequately" (e.g., Huang et al., 2014).[6] This intuition is likely linked to intuitions about consciousness and agency: how can a machine be to blame? However, in some cases, we do blame the functioning and design of machines: I may not know how to use my phone, but also, my phone's design or functioning may be faulty or suboptimal relative to my needs. We will argue that the latter argument applies to AI as well, against the common notion that the programmer is always the one to blame or the one biased, in Sect. 6. On the other hand, in the current subsection we unpack the idea that becoming a better programmer is a strategy that helps but does not solve the AP and hence needs to be aided by other strategies informed by our framework above.

The first set of strategies is preventing humans from "bending rules", i.e., to never depart from certain linguistic references. If these references are the ideals of formal languages, like mathematics and programming, then the chance of a command from a human to a machine leading to an unexpected result should diminish. In fact, this is what learning mathematics and programming is in principle about: we learn what strings of symbols entail, either within the symbolic realm (e.g., the iteration of $y = x^2$) or outside it (e.g., a sheet of paper coming out of a printer after commanding "print").

The stipulation that programmers should become "good" programmers from a formal standpoint is mainstream. Accordingly, there are many strategies in this direction. A key notion of the classical paradigm of logic, namely formal verification, is a major example of this. The key idea here is to change the practice of coding in a way that makes it capitalise on the power of logic. Programs are not checked anymore by testing applications but in a formal way, close to the ideals of logic. One proves (often with interactive theorem provers) that some properties hold in the form of real theorems (in the sense of theorems in mathematics). The theorem provers produce a formal *certificate* vouching for the correctness of the theorem. Similarly,

[6] For more informal reflections, see: https://towardsdatascience.com/dont-blame-the-ai-it-s-the-humans-who-are-biased-d01a3b876d58; https://blog.codinghorror.com/the-first-rule-of-programming-its-always-your-fault/; https://towardsdatascience.com/dont-trust-ai-10a7df520925

model checking is done currently in practice in the context of, e.g., autonomous vehicles (cf. Sifakis & Harel, 2023). In these approaches, formal properties of a program get verified by deduction (or something logically equivalent, like the construction of models), i.e. by a finite set of rules, showing that the assumptions made about the program guarantee a property of the outcome of the computations. In this way, programming can be seen as an exact science. As Hoare (1969) puts it, "Computer programs are mathematical expressions. They describe with unprecedented precision and in the most minute detail, and behaviour, intended or unintended, of the computer on which they are executed". One does not need to listen to the programmer's intentions to understand that an algorithm terminates or that an input of this sort will never yield an output of that sort. One can simply trust the certificates once a corresponding theorem is stated and proven. The step from the code to this mathematical expression—so to speak—is often connected to a strong emphasis on logics, as in endeavours like proof extraction (cf. Letouzey, 2008), where constructive proofs get turned into algorithms. There is also a step of abstraction involved here as the properties of programs are normally abstracted from the actual code. People have long been not just thinking about this on a technical level alone but also concerning protocols (cf. Rogers, 1981). Yet, current autonomous systems are too complex to be formally verified, and thus this strategy meets limitations (and AI researchers are aware of this).

However, besides this known limitation, our framework conceptualises a more substantial one. Note that we explicitly problematized the case of $y = x^2$ above: we follow this rule as a result of training with the rule, and what this training entails should be further unpacked. Training humans in formal languages is an approach that in fact works to a sizable extent, but not because we humans are able to abstract ourselves away from the parameters in the onion model above. Instead, it is because training in formal languages constitutes input to these parameters. In fact, users of formal languages like computer scientists or mathematicians also belong to specific subcultures that influence their formal practices, as the field of ethnomathematics shows both historically and contemporarily around the notion of mathematical culture (e.g., Asper, 2009; Katz, 2016; Larvor, 2016; Ju et al., 2016; Verran, 2001).

In other words, mathematics and similar languages like programming languages are enculturated.[7] Thus, deviations from a given formalist ideal may be attributed to different things besides a lack of training in a formal language. Even if a given culture intended to create a globally homogeneous linguistic group using a universal formal language, training in sub-cultures may lead to slightly different uses of such a formal language. For instance, the learning of mathematics has social aspects that leak into mathematics itself and therefore it seems safe to infer that they also leak into programming practices. There is room for choice in mathematics (Lakatos, 1976), mathematicians depart from formalisms in their uses of mathematics

[7] Recently the enculturation of mathematics has been formally precised by modelling via frame semantics how mathematicians learn to create new mental prototypes for mathematical situations (like within different types of proofs) and symbols during their training (cf. Carl et al., 2021; Fisseni et al., 2019; Fisseni et al., 2023).

(Pérez-Escobar, 2022; Pérez-Escobar & Sarikaya, 2022; Wittgenstein, 1976, 1978; Pérez-Escobar, 2023b), and mathematical concepts themselves are indeterminate (Bangu, 2023; Pérez-Escobar, 2023a; Tanswell, 2018; Zayton, 2022).

Given this, it is very feasible that different humans have different expectations on how a command should be followed, regardless of whether it is expressed in natural or formal languages. In fact, broadly speaking, it is well known that individual psychological differences make a difference in human–computer interactions including programming and code-review (e.g., Bishop-Clark, 1995; Whitley, 1996; Da Cunha & Greathead, 2007). More specifically for our purposes, it has been found that personality and cognitive differences influence not only the frequency of mistakes (here understood as deviations from computer formalisms) but also the type of mistakes made by human programmers independent of their proficiency level (Huang et al., 2014). The latter study also suggests that increased training and proficiency reduces the variability of mistakes but does not eliminate it, in consonance with the remarks above that training in a community fosters alignment but is not the only factor: psychological and other social and cultural factors remain.

In other words, training homogenises rule-following, but not completely. Taken together, this suggests that psychological, social, and cultural factors indeed influence our use of formal languages, and that "becoming a better programmer" is a strategy that helps but will not solve the fundamental issue. As Wittgenstein said, we cannot walk on "slippery ice where there's no friction" (Wittgenstein, 2009, p. 107), i.e., our use of language is in constant friction with psychological, social, and cultural factors. For this reason, we should find complements to this strategy to address the AP. Informed by our Wittgensteinian framework, Sect. 6 aims to do this.

The sceptic may propose other explanations for human deviations from the deterministic formalism. One such explanation is that fatigue and attention/arousal issues lead to mistakes which are not connected to human rule bending: the human knows, for instance, why their code is wrong and what they should have written so that the machine did otherwise, but they did not pay enough attention and made a mistake in their command. This is indeed a possibility, yet not incompatible with the above social aspects of (formal) rule following. Not only this, but such attention issues are sometimes gateways for the manifestation of the individual psychological factors of the onion model. People are not extremely meticulous constantly (e.g., constantly doubting oneself, rereading multiple times, etc.) but normal practices involve more "natural" behaviour on a daily basis (e.g., there are mathematical "hinges", unfalsifiable or epistemically resistant beliefs that are often implicit and constitute the bedrock of mathematical practices (Kusch, 2016; McGinn, 1989; Moyal-Sharrock, 2005; Wittgenstein, 1972); furthermore, mathematicians themselves skip formal levels of analysis on a daily basis, as Wagner (2022) observes). Sometimes, oversights tend to converge into a type of mistake, thus becoming systematic biases and revealing important psychological information about humans. When someone is engrossed in a particular line of thinking, attentional biases can account for their inability to contemplate other potential options (e.g., Baron, 2000). In the context of language, these biases are psychological manifestations on the approach to a problem using symbols: decreasing attention amplifies bias manifestation instead of being necessarily a factor outside the model. Of course, given our

framework above, it is unclear whether rule bending should be qualified as a source of bias at all, since it would assume a fixed reference.

The sceptic may propose another challenge. The AP manifests itself when we program a command and the machine executes an unexpected response. If I am biased and the machine reproduces my bias according to my intentions, there is no AP. This is correct, but it is not the case of bias we are concerned with in this work. Instead, we are concerned with "biases" in the use of language (to talk about bias in this sense, we need a reference: how a machine interprets a formal language; thus, we are concerned with "biases" as departures from this reference). These "alignment biases" in the use of language lead to unexpected responses by machines. Thus, for instance, we are concerned with those instances of algorithmic biases where the human did not intend the machine to reflect such biases. Human bias (explicit or implicit) and the AP (human "bias" as deviation from machine use of language) are different problems that require different solutions. However, we must concede that these problems may be hard to demarcate in practice. For instance, ChatGPT commits to specific political positions in a way qualified as algorithmic bias (e.g., McGee, 2023; Motoki et al., 2023; Rozado, 2023; Rutinowski et al., 2023), but this may be the intentional outcome of human political bias *or* the result of misalignment.

All in all, humans, natural rule benders, may stop bending rules by focusing on "clear" specifications of human goals. This may be achieved, for instance, by training in programming and computer science, in the formal aspects of "unbent" machine language. However, this training and linguistic clarifications are patchworks; they do not fully address the underlying principle at the base of the problem. In the context of future powerful AIs, mistakes like these can be fatal.

6 Aligning from the Side of Machines

As discussed before, "proper rule-following" is underdefined by rules, and while clarification may help, this issue is in principle unsolvable. To keep people aligned, they need to share psychological, social, and cultural factors, like common training in a linguistic practice. According to Wittgenstein's notion of "meaning as use", the best way to understand a rule is to see how it is used in natural contexts. In *Philosophical Investigations*, Wittgenstein claims that "Obeying a rule is a practice" (Wittgenstein, 2009, paragraph 202). For instance, to understand the command "bring me food", the best that one can do is to see how people follow the command in a given context. We believe that AI training procedures can be inspired by this consideration and the broader framework presented in Sect. 3, thus complementing the strategy presented in Sect. 5 for the development of safe AI.

First, let us briefly introduce two very common AI training procedures: exception-adding and desiderata-listing. Exception-adding consists in restricting undesirable outcomes produced by the AI whenever (or before) they appear. For instance, if we instruct an AI to store oranges in a 3D space as efficiently as possible, the AI may squeeze the oranges to maximise the number of stored oranges. If squeezing the oranges is undesirable (for instance, because this way they cannot be consumed after

transportation) we stipulate squeezing as an exception, and then the machine would find the next best path of action. A real example that became mediatic was an object/person classifier by Google that mislabelled people of colour as "Gorillas"; the labelling tool was forbidden to tag anything as "Gorilla", "Monkey", "Chimp" etc. (cf. Simonite, 2018; Wilner, 2018). Desiderata-listing consists in listing all valuable parameters so that the machine finds solutions that do not compromise on them, or compromises on them only relative to their order in a hierarchy. For instance, in the example above, desiderata-listing would consist in listing not just storing as many oranges as possible, but also keeping the orange juice within the orange skin, among other desiderata. Unlisted desiderata will often be compromised. An intuitive way to understand the latter strategy is via the wish-granting genie in a lamp meme: while the genie grants our wish, the way that the wish is realised unexpectedly compromises on some implicit desideratum. For instance, someone tells the genie "I wish to be famous", and the genie kills the person, followed by a final vignette depicting a newspaper with the heading "man murdered by real genie!".[8] Proper desiderata-listing should have included "I must live" besides being famous.

Not only language underdetermines "proper rule-following", but also, different people may expect different executions of rules across different scenarios. Programmers, as individuals or groups, often generalise their values concerning the parameters of the onion model too far. A systematic effort to counter overgeneralization (though not directly informed by the Wittgensteinian framework presented here) has been done. This includes diversity issues in teams that can hinder the generalizability of their output to some extent, as AI development does not happen in a vacuum but in the real environments of company locations, often in the US. The framework here presented can inform and enrich such efforts. Note that such a diversity is not relevant here for the purposes of social justice initiatives, but to improve on the AP by adding design and training variety to AIs (this connects to other debates in philosophy; see, e.g., Longino, 1993 and Foley, 2003).

Data collection can also benefit from this approach. A prominent example is the "Moral Machine experiment" (Awad et al., 2018), an ambitious global study initiated by the MIT to understand human preferences in the context of moral dilemmas faced by autonomous vehicles. People around the world have given their opinion on what a self-driving car should do in certain scenarios. In all scenarios, a collision was unavoidable, but depending on the action taken the outcomes differ. For instance, the car can either compromise the safety of young passengers in a car or elderly pedestrians. Shall the car prioritise the pedestrians or the safety of passengers? Should one take into account in the moral judgement whether the pedestrians were jaywalking? The study by Award et al. (2018) shows that there are cultural differences concerning moral judgements in this scenario. While some ways

[8] For the original comic strip, see: https://explosm.net/comics/kris-desire. The genie analogy seems to appear for the first time in the following interview in 2014: https://www.edge.org/conversation/the-myth-of-ai. It is claimed that "This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want." Yet, this judgement assumes that machine use of language is a reference for linguistic orthodoxy, and hence, one really asked for what the machine did.

of proceeding are more agreed upon, like sparing humans over animals, there is disagreement as for others. For instance, the paper observes that "countries belonging to the Southern cluster show a strong preference for sparing females compared to countries in other clusters." It is results like this that have inspired the need to add (in this case, cultural) diversity in AI development teams.[9] This entails that, given generic programming rules, different people may expect an automated car to do something different. More specifically, according to our framework, these differences may manifest as linguistic and rule-following differences relevant for the context of developing automated cars, at least sometimes in ways too implicit to be accounted for by standard procedures like exception-adding and desiderata listing. Remember that, according to our framework, different people may expect different executions of a rule even when it is surrounded by substantial clarification, as in the form of desiderata and exceptions. Diversity with regards to the stipulation of exceptions and desiderata is, therefore, not enough.

Exception-adding and desiderata-listing are also relevant in the interplay of expert systems with deep learning systems if we assume the desiderata to be previously hard coded and not extracted from the learning set directly. However, as we are about to see, the learning set by itself presents another angle on how to achieve better aligned systems.

Exception-adding and desiderata-listing have problems that become evident after the presentation of our framework in Sect. 3. Not only may we lack the resources to list all the exceptions and desiderata for a given rule across contexts[10] (and thus exhausting in this way how a rule ought to be followed is problematic) but also there is variability in how all this is linguistically presented relative to psychological, social, and cultural factors. To improve on these weaknesses, we propose that AI training procedures should provide a synthesis of 1) commands, 2) their executions across contexts (where their true meaning resides), and 3) across different psychological, social, and cultural values that underlie the relationship between 1 and 2. This can be achieved in at least three ways: by vicarious learning, by feedback training, and by software diversity, all integrating 1, 2 and 3.

Vicarious learning procedures should consist of presentations of commands and their executions by two or more aligned agents across multiple contexts in a way that the machine is sensitive to the psychological, social, and cultural parameters of the involved agents.[11] This way, the "correct" manner of following a rule is not exhausted by exceptions and desiderata, but shown in practice across different contexts (the machine would translate the presentations into formal rules of its own, but this is free from the AP). Furthermore, the machine not only pairs commands and

[9] Another case of misbehaving AIs is the misclassification of people with darker skin tones in many visual classifiers (cf. Crawford, 2016).

[10] "It is certainly very hard, and perhaps impossible, for mere humans to anticipate and rule out in advance all the disastrous ways the machine could choose to achieve a specified objective" (cf Russell, 2019).

[11] It is convenient to note here that, given our framework and that we are applying and adapting the literature on the later Wittgenstein and AI, vicarious learning here is based on the analysis of overt performance, and not on empathy, the observation and acquisition of values, and similar.

their executions across contexts, but also associates the psychological, social, and cultural parameters of the agents to those pairs. Moreover, different training agents encompassing different values of psychological, social, and cultural parameters should be employed. This way, fluctuations in these parameters and contexts, which is reflected in different executions of commands, are accounted for.

Feedback training can also synthesise commands, executions across contexts and psychological, social, and cultural factors. It should consist of executions of commands by the machine across contexts, but instead of adding exceptions after undesirable executions, the trainers only tell the machine whether the execution is satisfactory or not. The trainer does not explain why (as this would introduce the AP again), but instead, the machine pairs the psychological, social, and cultural values of the trainers to their feedback (satisfactory or not) on the execution for each context.

The third procedure is to foster software diversity. The main difference with the other two is that there is not a single software that is made sensitive to psychological, social, and cultural factors, but instead there are multiple software tailored to specific combinations of values of these parameters. As mentioned before, it has been experimentally shown that people with different psychological profiles make different programming errors. Very importantly, the same study also shows that software diversity reduces the frequency of such errors (Huang et al., 2014). According to our framework, these errors are derived from the AP, despite the authors calling these errors "human errors" in line with the trend of blaming the human side of the alignment: psychological factors are one of the keys to human alignment. Indeed, from this perspective, software diversity improves alignment because some software are a better fit for a given psychological value of the parameter in the above onion model than others.

Thus, a third strategy to improve alignment from the side of AI design and training is to do as above (vicarious learning and feedback training) in a way that commands and their executions across contexts are synthesised, but without adding psychological, social, and cultural factors – instead, different software (e.g., AI models) would specialise in rule-following under specific value combinations of the parameters of the onion model. While this specificity may improve on alignment precision, there is the risk that software may be employed in suboptimal ranges of the parameter values. A very simple example (which by virtue of its simplicity is seldom problematic in practice) in natural language is the cultural factor of being a native Spanish speaker and handling double negations similarly to simple negations in English: not accounting for this in natural language processing would result in misalignment. This approach deviates from the one-solution-fits-all approach to AI, the latter being especially relevant as current ground models are only offered by a handful of companies.

7 A Limitation of both Strategies: How Should a rule be Applied in a new Situation?

In this section we discuss an issue that none of the strategies above can solve, although awareness about it may help to some degree. Let us recall that a bias is a departure from a normative standard. Hence, programming errors may be understood as a bias if compared to machine use of language. Similarly, the general notion of algorithmic bias is understood as departures from moral, legal, statistical or other kinds of normative standards (Antony, 2016; Danks & London, 2017; Fazelpour & Danks, 2021; Johnson, 2021). Thus, in a way similar to our approach, Fazelpour and Danks (2021) state that:

> algorithmic bias is not a function solely of the code or mathematics, but also depends on the domains of application, goals for the algorithm use, and other contextual factors." (Fazelpour & Danks, 2021, p. 2)

Given that algorithmic bias depends on normative standards, domains of application, and contextual factors (all similar to the idea that rules are followed differently across contexts), it is not clear how a rule ought to be followed, or a command executed, in an unforeseen context. After all, command execution under different contexts is one of the factors synthesised in the strategies depicted in Section 6. In unforeseen contexts, there is no normative standard against which bias, misapplication, misexecution, or bad use can be defined. Again, the symbols that comprise natural and formal languages do not exhaust their uses, not only because of variability across psychological, social, and cultural factors, but also because of contextual variability, and the case of new contexts is especially delicate.

Note that this is not a problem exclusive to the AP. In fact, new contexts are a common source of disagreement and misalignment between humans (in an amplified manner if the humans also feature different values of the parameters of the onion model) which makes it harder to locate a "bias". To recapitulate: while different values of psychological, social, and cultural parameters are a source of misalignment, so is the advent of new, unforeseen contexts, and the latter cannot be fully accounted for by the alignment strategies in Sects. 5 and 6 (and perhaps any strategy whatsoever).

The following are some examples of the underdetermination of rules in new contexts in different areas. They illustrate the problems and difficulties of the application of rules in new contexts of which we should be aware: if we aim to at least improve on the AP between humans and machines, we must acknowledge how alignment between humans is itself problematic in these situations.

- In chess: chess is a game with simple rules often used as a toy example to illustrate a formalist view of mathematics: once the rules are established, they determine what is possible in the system with clarity. However, imagine the following unanticipated scenario: in a chess tournament where each player has one minute to make a move or else they lose their turn, a player accidentally knocks

some figures out of their place in their turn, and proceeds to put each figure in its place. Does the time used to fix this accidental situation count towards the limit? Of course, post hoc, new rules can be stipulated to further clarify what to do in such a situation, but the point is that the new rule would have loopholes of its own, which may become apparent again in future contexts.

- In law: existing legal codes are unprepared to handle phenomena from the current digital era. For instance, it was controversial whether Apple's full control of the App Store in their iPhones constituted a monopoly or not, given that the iPhone as a product faces competition. Society has mechanisms to try to reach consensus among divergent opinions, like courts, not only to balance the different interests of the parties involved, but also due to the indeterminacy of rule-following especially in new contexts (which the involved parties often exploit in their favour). In Common Law systems, previous rulings on (then) new cases are used as references for future cases (similarly to the post hoc stipulation of new rules in the case of chess above).
- In mathematics: new mathematical entities stirred the mathematical landscape in Early Modern Europe because the rules that worked for the old entities could not control the new ones in the same way, leading to controversies and disagreements. Examples include negative numbers (for which some rules of proportion do not work and hence it was not clear if they are numbers at all, as Antoine Arnauld showed (Heeffer, 2011, p. 867)), infinite (for which it was not clear whether and how "the whole is bigger than the parts" worked), imaginary numbers (for which representations in the geometry of the time did not seem to work, which in turn questioned the validity of methods that employed them) and 0 (for which it was not clear whether it could divide another number). Again, like in the cases above, new axiomatic systems were defined *after* the encounters with the new systems, involving processes of resolution of different characteristics and lengths, to regulate mathematical rule-following in the new landscape. The encounters with incommensurability in Ancient Greece is another classical example of this phenomenon; see Friedman (2024) for a recent discussion of the perceived lack of control of mathematicians during these situations.

Wittgenstein emphasised that rule-following should be seen performed in practice:

> Not only rules, but also examples are needed for establishing a practice. Our rules leave loopholes open, and the practice has to speak for itself. (On Certainty 139; Wittgenstein, 1972)

The issue with unanticipated contexts is precisely that we have never been exposed to how rules are followed in that context, and illustrates very well how natural or formal language alone, without examples of their uses in a variety of contexts, underdetermines rule-following. Where practical examples of rule-following are lacking, which is always the case in unanticipated new contexts, the loopholes remain open, and disagreements and misalignment ensue.

Despite this, humans can reassess the situation ad hoc under new contexts according to their subjective assessment. For instance, humans may find inspiration

in how economic debt works to control negative numbers, or debate whether the chess player was able to strategize while fixing the accident. This is a resource that machines lack: machines do not "reassess", but process symbols deterministically (or in Shanker's words, again, they follow rules "mechanically"). They must be preloaded with all relevant guidelines (including stipulations on how to learn and "pseudoimprovise"), and thus are incapable of improvising in this human manner. In other words, language and some paradigmatic associated uses lead to background assumptions on the use of language, but these background assumptions are not enough to cover unanticipated contextual contingencies. This may lead to hesitation and disagreement in humans concerning how to proceed (how a rule should be "bent") and adjust to the new situation, while deterministic machines do not hesitate nor disagree. This shows how, in fact, rule bending is a necessary and positive phenomenon in human rule-following.[12] Another positive note is that, even during rule-following in new contexts, humans with similar training tend to behave in similar ways, as recent Wittgensteinian scholarship has discussed (Berg, 2024), thus staying aligned. This is why human–human alignment is often acceptable even in many new contexts.

These human ad hoc adjustments to new contexts, where misalignment is more common, are just as susceptible to be influenced by the parameters of the onion model. This ad hoc adjustment tries to find the best subjective application of a rule in the new context. Thus, the hope is that accounting for psychological, social, and cultural variability reduces misalignment with machines when new contexts are presented, as they do between humans, taming this AI alignment limitation to some extent. Incidentally, this may also address Shanker's initial worry about the lack of flexibility of computers, discussed early in Section 2. This will be the subject of future work.

8 Conclusion

The AP, as outlined at the beginning of this paper, poses a significant and urgent challenge to the rapid advancements of AI technology for many reasons. For instance, governments and authorities need to quickly adapt to a new technological landscape, and the AP makes this more challenging. We have argued that Wittgensteinian ideas on rule-following and our alignment model offer a framework that helps thinking about possible tools. The three dimensions of human alignment (individual, social, and cultural) should all be checked when developing alignment

[12] In fact, the notion of emergent bias, consisting in the application of algorithms in new contexts for which they were not devised (e.g., Friedman and Nissenbaum, 1996; Mann and Matzner, 2019), depends on this ad hoc assessment to be conceptually sound: there cannot be a bias in new contexts without a normative reference, and thus, the ad hoc adjustment is adapted as the normative reference (despite high levels of misalignment between humans themselves concerning such an assessment!). The exception to this is when AIs are trained under situations different from the real world situations where they are employed, despite these real world situations being known and anticipated by humans who are aligned in the application of a rule in these situations.

strategies. As we saw, existing efforts are trying to address this worry to some extent, but we have provided more specific, basic guidelines based on our framework that, we argue, will further aid in taming the AP.

Our journey through Wittgenstein's philosophy reminds us that there is no unambiguous essence residing in symbols or code. Their meaning emerges in their reception in practice, where linguistically indeterministic humans are deeply involved. By embedding AIs in similar dynamic, feedback-rich contexts maximising diversity among all three mentioned dimensions, and ensuring that they perform their own rule-following within structured societal frameworks, we may be able to influence their trajectory to align them closely to our own values and intents.

The strategies that we have proposed are by no means the silver bullet for the AP, but they are instead a Wittgensteinian philosophical enrichment (the first of this kind for AI safety in general, and the AP in particular, as far as we know) rather than strict alternatives to what AI researchers currently do. Furthermore, the philosophically-informed methods suggested here would benefit from experimental testing in further works. We stress that the helpfulness of our Wittgensteinian approach to the problem of AI safety underscore the need for an interdisciplinary dialogue, one where the humanities and technology are in continuous exchange. As AI continues its ascent, the value of such cross-disciplinary collaborations will only increase.

Acknowledgements We are grateful for the helpful clarificatory comments of two anonymous reviewers.

Authors' Contributions Both authors contributed equally to all aspects. JAP-E and DS: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing, Review, Editing. Both authors authors have read and approved the manuscript.

Funding Open access funding provided by University of Geneva. The first author has been supported by two Postdoc. Mobility project grants by the Swiss National Science Foundation (P500PH_202892; P5R5PH_214160). The second author is thankful for the financial and ideal support of the Studienstiftung des deutschen Volkes and the Claussen-Simon-Stiftung as well as the Research Foundation Flanders (FWO) [grant number FWOAL950]. The views stated here are not necessarily the views of the supporting organisations mentioned in this acknowledgement.

Data Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent for Publication Not applicable.

Competing Inteests The authors declare that they have no competing interests.

References

Andrus, M., Dean, S., Gilbert, T. K., Lambert, N., & Zick, T. (2021). AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks. Retrieved from https://arxiv.org/abs/2102.04255

Antony, L. (2016). Bias: Friend or foe? Reflections on Saulish skepticism. *Implicit Bias and Philosophy, 1*, 157–190.

Arnold, Z., & Toner, H. (2021). AI Accidents: An Emerging Threat. Retrieved from https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/

Asper, M. (2009). The two cultures of mathematics in Ancient Greece. In E. Robson & J. Stedall (Eds.), *The Oxford Handbook of the History of Mathematics* (pp. 107–132). Oxford: Oxford University Press.

Awad, E., Dsouza, S., Kim, R., et al. (2018). The Moral Machine experiment. *Nature, 563*, 59–64.

Bangu, S. (2023). Wittgenstein on Proof and Concept-Formation. *The Philosophical Quarterly*, pqad111.

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Berg, Á. (2024). Was Wittgenstein a radical conventionalist? *Synthese, 203*(2), 37.

Bishop-Clark, C. (1995). Cognitive style, personality, and computer programming. *Computers in Human Behavior, 11*(2), 241–260.

Carl, M., Cramer, M., Fisseni, B., Sarikaya, D., & Schröder, B. (2021). How to frame understanding in mathematics: A case study using extremal proofs. *Axiomathes, 31*(5), 649–676.

Casey, G. (1988). Artificial Intelligence and Wittgenstein. *Philosophical Studies, 32*, 156–175.

Chrabaszcz, P., Loshchilov, I., & Hutter, F. (2018). Back to basics: Benchmarking canonical evolution strategies for playing atari. *arXiv preprint arXiv:1802.08842*. Retreived from https://arxiv.org/abs/1802.08842

Cook, J. (2018). Amazon scraps "sexist AI" recruiting tool that showed bias against women. *The Telegraph, 10*, 10.

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York times, 25*(06), 5.

Da Cunha, A. D., & Greathead, D. (2007). Does personality matter? An analysis of code-review ability. *Communications of the ACM, 50*(5), 109–112.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In C. Sierra (Ed.), *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 4691–4697). Melbourne: AAAI Press.

Dreyfus, H. (1978). *What computers can't do: The limits of artificial reason*. Harper & Row.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass, 16*(8), e12760.

Fisseni, B., Sarikaya, D., Schmitt, M., & Schröder, B. (2019). *How to frame a mathematician: Modelling the cognitive background of proofs* (pp. 417–436). Univalent Foundations, Set Theory and General Thoughts.

Fisseni, B., Sarikaya, D., & Schröder, B. (2023). How to frame innovation in mathematics. *Synthese, 202*(4), 108.

Foley, D. (2003). Indigenous epistemology and Indigenous standpoint theory. *Social Alternatives, 22*(1), 44–52.

Friedman, M. (2024). On metaphors of mathematics: Between Blumenberg's nonconceptuality and Grothendieck's waves. *Synthese, 203*(5), 1–27.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS), 14*(3), 330–347.

Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., ... & Levy, K. (2021). Artificial intelligence, systemic risks, and sustainability. *Technology in Society*, 67, 101741.

Gerrard, S. (1995). Wittgenstein Versus Artificial Intelligence?. In K. Gavroglu, J. Stachel and M. W. Wartofsky (eds.), *Science, Mind and Art: Essays on science and the humanistic understanding in art, epistemology, religion and ethics In honor of Robert S. Cohen* (pp. 89–98). Dordrecht: Springer Netherlands

Hao, K. (2019). AI is sending people to jail—And getting it wrong. *MIT Technology Review*. https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

Harre, R. (1988). Wittgenstein and artificial intelligence. *Philosophical Psychology, 1*(1), 105–115.

Heeffer, A. (2011). Historical objections against the number line. *Science & Education, 20*, 863–880.

Hoare, C. A. R. (1969). An axiomatic basis for computer programming. *Communications of the ACM, 12*(1969), 576–580.
Huang, F., Liu, B., Song, Y., & Keyal, S. (2014). The links between human error diversity and software diversity: Implications for fault diversity seeking. *Science of Computer Programming, 89*, 350–373.
Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese, 198*(10), 9941–9961.
Ju, Löwe, Müller, & Xie (eds.) (2016). *Cultures of Mathematics and Logic*. Cham: Springer Birkhäuser.
Katz, V. (2016). The mathematical cultures of medieval Europe. In L. Radford, F. Furinghetti, & T. Hausberger (eds.), *Proceedings of the 2016 ICME Satellite Meeting of the International Study Group on the Relations Between the History and Pedagogy of Mathematics* (pp. 39–64). Montpellier, France: IREM de Montpellier.
Kusch, M. (2016). Wittgenstein on mathematics and certainties. *International Journal for the Study of Skepticism, 6*(2–3), 120–142.
Larvor, B. (2016). What are mathematical cultures? In S. Ju, B. Löwe, T. Müller, Y. Xie (eds.), *Cultures of Mathematics and Logic* (pp. 1–22). Cham: Springer Birkhäuser.
Letouzey, P. (2008). Extraction in coq: An overview. In A. Beckmann, C. Dimitracopoulos and B. Löwe (eds.), *Logic and Theory of Algorithms: 4th Conference on Computability in Europe, CiE 2008, Athens, Greece, June 15–20, 2008 Proceedings 4* (pp. 359–369). Berlin and Heidelberg: Springer.
Longino, H. (1993). Feminist standpoint theory and the problems of knowledge. *Signs: Journal of Women in Culture and Society*, *19*(1), 201–212.
Mann, M., & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society, 6*, 2.
McGee, R. W. (2023). Is Chat Gpt Biased Against Conservatives? An Empirical Study. *Working Paper*. Available at https://ssrn.com/abstract=4359405
McGinn, M. (1989). *Sense and Certainty: A Dissolution of Scepticism*. Blackwell.
Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More Human than Human: Measuring ChatGPT Political Bias. *Working Paper*. Available at https://ssrn.com/abstract=4372349
Moyal-Sharrock, D. (2005). *Understanding Wittgenstein's On Certainty*. Basingstoke: Palgrave.
Obermeier, K. K. (1983). Wittgenstein on language and artificial intelligence: The Chinese-room thought experiment revisited. *Synthese, 56*(3), 339–349.
Pérez-Escobar, J. A. (2022). Showing mathematical flies the way out of foundational bottles: the later Wittgenstein as a forerunner of Lakatos and the philosophy of mathematical practice. *KRITERION-Journal of Philosophy, 36*(2), 157–178.
Pérez-Escobar, J. A. (2023a). A new role of mathematics in science: Measurement normativity. *Measurement, 223*, 113631.
Pérez-Escobar, J. A. (2023b). The role of pragmatic considerations during mathematical derivation in the applicability of mathematics. *Philosophical Investigations*. https://doi.org/10.1111/phin.12412
Pérez-Escobar, J. A., & Sarikaya, D. (2022). Purifying applied mathematics and applying pure mathematics: How a late Wittgensteinian perspective sheds light onto the dichotomy. *European Journal for Philosophy of Science, 12*(1), 1.
Rogers, R.(1981). Planning for independent software verification and validation, *AIAA 1981–2100. 3rd Computers in Aerospace Conference*.
Rozado, D. (2023). The political biases of chatgpt. *Social Sciences, 12*(3), 148.
Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.
Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., & Pauly, M. (2023). The Self-Perception and Political Biases of ChatGPT. *arXiv preprint* arXiv:2304.07333. Retreived from https://arxiv.org/abs/2304.07333
Shanker, S. G. (1998). *Wittgenstein's Remarks on the Foundations of AI*. Routledge.
Sifakis, J., & Harel, D. (2023). Trustworthy Autonomous System Development. *ACM Transactions on Embedded Computing Systems, 22*(3), 1–24.
Simonite, T. (2018). When it comes to gorillas, Google photos remains blind. *Wired*. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/
Tanswell, F. S. (2018). Conceptual engineering for mathematical concepts. *Inquiry, 61*(8), 881–913.
Verran, H. (2001). *Science and an African logic*. University of Chicago Press.
Wagner, R. (2022). Mathematical consensus: A research program. *Axiomathes, 32*(Suppl 3), 1185–1204.
Whitley, B. E., Jr. (1996). The relationship of psychological type to computer aptitude, attitudes, and behavior. *Computers in Human Behavior, 12*(3), 389–406.

Wiener, N. (1960). Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science, 131*(3410), 1355–1358.
Wilner, A. S. (2018). Cybersecurity and its discontents: Artificial intelligence, the Internet of Things, and digital misinformation. *International Journal, 73*(2), 308–316.
Wittgenstein, L. (1972). *On Certainty*. Harper & Row.
Wittgenstein, L. (2009). *Philosophical Investigations, Revised* (4th ed.). Wiley-Blackwell.
Wittgenstein, L. (1976). *Wittgenstein's Lectures on the Foundations of Mathematics*, C. Diamond (Ed.). Ithaca: Cornell University Press.
Wittgenstein, L. (1978). *Remarks on the foundations of mathematics* (3rd revised edition), G. H. von Wright, G. E. M. Anscombe and R. Rhees (Eds.), G. E. M. Anscombe (Trans.). Oxford: Basil Blackwell. First edition published in 1956.
Yingjin, X. U. (2016). Does Wittgenstein Actually Undermine the Foundation of Artificial Intelligence? *Frontiers of Philosophy in China, 11*(1), 3–20.
Zayton, B. (2022). Open texture, rigor, and proof. *Synthese, 200*(4), 341.
Zhang, T., Rashidinejad, P., Jiao, J., Tian, Y., Gonzalez, J. E., & Russell, S. (2021). MADE: Exploration via Maximizing Deviation from Explored Regions. Retrieved from https://proceedings.neurips.cc/paper/2021/hash/5011bf6d8a37692913fce3a15a51f070-Abstract.html

Authors and Affiliations

José Antonio Pérez-Escobar[1,2,3] · Deniz Sarikaya[4,5]

✉ José Antonio Pérez-Escobar
jose.perezescobar@unige.ch

✉ Deniz Sarikaya
deniz.sarikaya@vub.be; deniz.sarikaya@uni-luebeck.de

[1] Centre Cavaillès, UAR 3608 République Des Savoirs, École Normale Supérieure, PSL University, Paris, France

[2] Department of Logic, History and Philosophy of Science, UNED, Madrid, Spain

[3] Department of Philosophy, University of Geneva, Geneva, Switzerland

[4] Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel (VUB), Brussels, Belgium

[5] Ethical Innovation Hub, Universität zu Lübeck, Lübeck, Germany