AI and Ethics
https://doi.org/10.1007/s43681-024-00496-5

ORIGINAL RESEARCH

Aligning artificial intelligence with moral intuitions: an intuitionist approach to the alignment problem

Dario Cecchini[1] · Michael Pflanzer[1] · Veljko Dubljević[1]

Received: 25 February 2024 / Accepted: 14 May 2024

Abstract

As artificial intelligence (AI) continues to advance, one key challenge is ensuring that AI aligns with certain values. However, in the current diverse and democratic society, reaching a normative consensus is complex. This paper delves into the methodological aspect of how AI ethicists can effectively determine which values AI should uphold. After reviewing the most influential methodologies, we detail an *intuitionist* research agenda that offers guidelines for aligning AI applications with a limited set of reliable moral intuitions, each underlying a refined cooperative view of AI. We discuss appropriate epistemic tools for collecting, filtering, and justifying moral intuitions with the aim of reducing cognitive and social biases. The proposed methodology facilitates a large collective participation in AI alignment, while ensuring the reliability of the considered moral judgments.

Keywords Alignment problem · Moral intuitions · Value pluralism · Hybrid approach

1 Introduction

The development and growth of artificial intelligence (AI) in recent years have spurred important challenges for humanity. A crucial issue is to ensure that AI systems benefit society, while helping realize human values. This question, called the *alignment problem*, has been extensively discussed in recent years [11, 24, 49]. A particularly important subject of discussion concerns which values AI should align with and, prior to that, which procedure AI ethicists should implement to ascertain the relevant values. This essay addresses this latter methodological question.

It is debated whether, and in what way, AI stakeholders' individual preferences should contribute to AI value setting. On the one hand, if AI systems were aligned with the soundest set of principles deliberated by a selected group of experts (Cfr. [1, 22]), regardless of laypeople's moral beliefs, one would risk overlooking value pluralism in society and context-sensitive problems that emerge from practice. On the other hand, aligning AI with the values implied by the majority of stakeholders' preferences (Cfr. [46, 49]) might perpetuate existing biases in society. For these reasons, a "hybrid" methodology, combining the inclusion of stakeholders' beliefs with ethical reflection, is reaching consensus in the recent literature [51, 57, 58]. However, some challenges remain unaddressed. Specifically, what type of individual beliefs has to be targeted to understand stakeholders' values? What kind of research strategies have to be employed to mitigate bias? This paper details an *intuitionist* research agenda that integrate hybrid approaches by tackling these questions. Our proposed methodology offers guidelines for aligning AI applications with a limited set of reliable moral intuitions, each underlying a refined cooperative view of AI. As we will argue, this method facilitates large-scale collective participation in AI alignment, while ensuring the reliability of the considered moral judgments.

We proceed as follows. Section 2 introduces the normative challenge raised by the alignment problem, distinguishing between *value setting* and *value implementation*. Then, Sect. 3 presents an overview of methodological approaches to AI value setting, from the limitations of purely top–down and bottom–up approaches to the more recent hybrid methods. Section 4 spells out our hybrid intuitionist approach. After defining moral intuitions as automatic and strong moral judgments, we discuss experimental methods for collecting and filtering them to mitigate cognitive and social

✉ Veljko Dubljević
veljko_dubljevic@ncsu.edu

[1] Department of Philosophy and Religious Studies, North Carolina State University, 101 Lampe Drive, Raleigh, NC 27695, USA

Published online: 27 May 2024

biases. Subsequently, we highlight how the inter-subjective process of normative justification can transform implicit moral content within intuitions into articulated cooperative models of AI applications. This stage would not only further correct biases but also facilitate the aggregation of intuitions into implementable goals for AI. To denote value disagreement that persists in the mitigation of biases and the process of normative justification, we introduce the concept of *reasonable intuition conflict*, which would be subject to public discussion and political deliberation. Finally, Sect. 5 discusses the main advantages of the intuitionist approach and addresses salient objections and limitations.

2 The alignment problem and value setting

While the path to a general AI is still distant, the present era is marked by a proliferation of *narrow* AI systems, programmed to accomplish specific tasks [21]. Some popular examples are virtual assistants, such as Amazon's *Alexa*, driving automation systems like Google's *Waymo*, and large language models like *ChatGPT*, released by OpenAI. Such systems are characterized by an extended operational autonomy, that is, the ability to act for extended time without a human operator. Furthermore, current AI technologies display a significant adaptability to the environment, which often translates into an increasing efficacy in delivering tasks.[1]

Delegating a growing number of tasks to autonomous artificial agents can potentially solve social problems and redirect human energies into desirable activities. However, the benefits of AI are contingent upon the objectives that the systems accomplish and possibly offset by the consequences they produce. In order to benefit society, AI systems must be directed toward the realization of certain defined values. However, given the extended autonomy and adaptability of AI, some systems may develop features that were not intended or foreseen by human designers ([57], p. 286). Therefore, the challenge is to design machines that mitigate social and environmental problems without introducing unacceptable harms or amplifying existing ones. In other words, humans must take control of the impact of AI on society. This general issue for AI has been defined by some authors as the *AI alignment problem* [11, 24, 49].

Although interpretations can differ, AI alignment is typically associated with the general concept of "beneficial AI" in the literature [25]. Admittedly, the notion of "beneficial AI" is vague, and its precise definition requires an independent essay. For the sake of the present discussion, we assume that an AI system is beneficial (i.e., aligned) whenever it contributes to human flourishing, encompassing physical health, individual happiness, and social well-being [52]. This entails that AI alignment is not understood here only as a problem of safety but also more broadly as a matter of regulating the consequences of AI on society (Cfr. [11, 31]).[2] Accordingly, examples of misaligned AI include not only the use of AI applications for criminal purposes ([21], pp. 113–141) but also large language models spreading misinformation [17], or data-driven algorithms that discriminate, thus perpetuating social inequalities [11, 42].

We identify two main phases in the process of value alignment.[3] First, one must address the normative challenge of determining the goals of the alignment, that is, *what values* AI systems should align with to be beneficial. We call this task *value setting*. Second, one must implement the identified goals into the AI systems, checking on their realization. This phase, which we define as *value implementation*, consists of understanding *how* AI should be designed to align with explicit values. The relevant challenge in this phase is to encode normative values using formal AI programming methods.[4] In this paper, we primarily focus on the value-setting aspect of the alignment problem while reserving the value implementation stage in the background as a necessary step in the process of aligning AI. Arguably, even in the value-setting stage, the *implementability* or *applicability* of the discussed values are important requirements when considering the ultimate practical goal of value alignment.

Setting the alignment goals requires establishing a clear hierarchy of human values to implement into AI applications. However, this requirement appears to contrast with multiple, sometimes competing, interpretations of beneficial AI in society. Certainly, persistent value disagreement within and across cultures makes AI value setting particularly challenging. For example, distinct ethical standpoints may result in different prioritizations of values [60]. *Virtue-centered* moralities might lean toward developing AI applications oriented toward realizing a common good and fostering positive relationships between citizens. On the other hand, *rights-centered* moralities may favor the development of AI

[1] For example, the ultimate version of ChatGPT (GPT-4) performs better than average in many academic and professional exams [43].

[2] Jonker [31] calls this aspect "social alignment", while distinguishing it from "value alignment", which concerns the safety of AI. By contrast, we understand "value alignment" more broadly, comprising social alignment.

[3] In a similar vein, Morley and colleagues [39] distinguish two aspects in AI ethics: the "what", i.e., the ethical principles for good AI, and the "how", i.e., the identification of the tools and methods to apply in the principles. Also, Gabriel [24] discerns the "technical" and "normative" aspects of value alignment and examine the connections between the two.

[4] The alignment process is likely to be iterative [57]. Following value implementation, developers receive feedback from the use of the systems. This feedback may prompt a recalibration of value setting.

applications that enhance individual freedom, property, and well-being.

Another reason for uncertainty about value setting is the existence of stakeholders with competing interests in AI. A notable example is the social dilemma in automated vehicles concerning the choice between protecting the vehicles' passengers and prioritizing the safety of vulnerable road users, such as pedestrians or cyclists [8]. Governments, for example, may be interested in protecting the most vulnerable users for public safety reasons, whereas private vehicle manufacturers have a financial incentive to prioritize their customers' safety.

As illustrated by such examples, the uncertainty about value setting underscores the need to establish a reliable methodology for filtering and selecting values that can be set as implementable goals for AI. For our context, reliability can be understood as the extent to which a certain method is conducive to beneficial AI systems (as previously defined).[5] In addition, we premise that the methodology must be suitable for informing democratic political institutions empowered to regulate AI. Therefore, beyond the perspective of beneficial AI, further constraints to the present discussion come from the normative boundaries of liberal democracy. Notably, democracy, among other things, prescribes respect for pluralism and human rights. We will consider these democratic requirements in evaluating a methodology for AI.

Granted these preliminary assumptions, the challenge is to identify a procedure inclusive enough to consider a wide range of evaluative viewpoints in society so that future AI does not discriminate or impose restrictive values. Nevertheless, such a procedure should also be able to integrate multiple social values into a set of implementable objectives for AI [51, 58]. The remainder of the paper aims to understand what kind of method fits the bill.

3 Methodological approaches to AI value setting: a brief overview

One plausible approach to the alignment problem is to start with a sound moral theory or a set of principles and then find the most appropriate tools and formal methods to apply them to AI systems. According to this approach, a moral inquiry should be conducted by a selected group of experts, say, for example, a scientific committee, which is empowered to deliberate a comprehensive set of values with which AI must align (Cfr. [1, 22]). Such a methodology, which we label as *top–down* (from principles to practice), dominated the AI ethics scene until five years ago and culminated in a proliferation of ethical guidelines for AI all over the world [30].

Admittedly, top–down approaches have contributed to delineating some universally shared principles for regulating AI.[6] Nevertheless, principle-based methodologies are prone to well-known and discussed criticisms. The primary issue lies in prioritizing and operationalizing principles in the AI practice [38]. Although some data reveal a convergence of ethical guidelines around fundamental principles (e.g., justice, privacy, beneficence, etc.), divergences arise regarding the interpretation of principles and how to resolve value conflicts emerging from practice ([30], p. 396).[7] Moreover, to be politically legitimate, interpretations of principles and trade-offs need to align with stakeholders' individual preferences in addition to the input of a group of experts whose epistemic authority is difficult to define given the elusive nature of AI ethics. The legitimacy of normative goals is crucial for the relationship of trust that needs to be established between AI agents and their stakeholders. This latter group may not trust machines that serve goals not aligned with their personal moral preferences. Therefore, the risk of neglecting individual evaluative beliefs is that potential users may opt out of using AI, thus nullifying all their expected benefits ([7], p. 110).

Motivated by these considerations, some authors have advocated for *bottom–up* methodologies, which aim to infer values from stakeholders' individual preferences (Cfr. [46, 49]). Rather than aligning with ethical principles, bottom–up approaches interpret beneficial AI as a system that best satisfies stakeholders' preferences. This approach enhances trustworthiness, democratic participation, and legitimacy in AI alignment. However, purely bottom–up methodologies have limitations concerning the aggregation and harmonization of individual beliefs. Specifically, aligning AI with the values implied by the majority of preferences risks producing a "tyranny of big data" and exploitation of minorities ([51], p. 655), which might be inconsistent with liberal democracy. Additionally, the quality of individual preferences, not just the quantity, seems to matter, assuming that not every belief has the same level of rationality.

The strengths and limitations of either a top–down or bottom–up approach point to the need for a *hybrid* method that combines top–down and bottom–up aspects according to the different demands of the alignment process [58]; that is, on the one hand, the need to consider a wide range of

[5] This means that the reliability of a methodology can be ultimately assessed by the long term consequences produced by AI on society. In the meanwhile, philosophers can debate about that based on rational expectations and predictions.

[6] Indeed, universal principles influenced the enactment of the first laws about AI in EU [18] and US [55].

[7] For example, the need to expand datasets to program fair, unbiased algorithms may conflict with individual privacy rights over personal information.

values from society and, on the other, the necessity to synthesize these values into implementable objectives for AI. Two recent accounts seem to align with this direction and are worth mentioning. The first one is the hybrid approach proposed by Umbrello and van de Poel [57] based on Value Sensitive Design (VSD). Specifically, the authors delineate a four-stage iterative design process to align AI technologies with social values. The agenda starts with an analysis of the values and needs of stakeholders, which are subsequently synthesized and translated into design requirements by relevant experts.

Similarly, Savulescu et al. [51] propose an approach called Collective Reflective Equilibrium in Practice (CREP), which includes data on public attitudes as input into a deliberative process aimed at determining AI policies. To ensure legitimacy and rational justification at the same time, Savulescu and colleagues argue that stakeholders' moral preferences have to be scrutinized for bias and prejudice; subsequently, policymakers have to seek an "overlapping consensus" between public attitudes and major ethical theories.

While VSD and CREP have significantly contributed to integrating laypeople's preferences, expertise, and ethical principles, they both fall short in defining the specific type of individual beliefs that must be targeted to understand stakeholders' values. Additionally, while both approaches recognize the need to filter individual preferences to inform AI value setting, they lack details on how AI ethicists can mitigate bias and select high quality data. Our intuitionist approach aims to advance the hybrid methodology by addressing these two fundamental aspects.

4 A hybrid intuitionist approach

In what follows, we spell out our intuitionist approach to AI alignment. The lesson derived from the previous discussion is that value setting should be sufficiently inclusive to involve evaluative beliefs from every potential AI stakeholder while screening and filtering those beliefs by using appropriate scientific tools and expertise to obtain reliable outputs. Thus, we define a hybrid *intuitionist approach* that fulfills such conditions. First, we outline our account of moral intuitions and explain how they can be appropriately collected and filtered according to some promising debiasing strategies (Sect. 4.1). Then, we show how justification can transform implicit moral content within intuitions into articulated cooperative views of AI (Sect. 4.2).

4.1 Reliable moral intuitions

Despite the highlighted limitations of the bottom–up approach, we contend that collecting individual evaluative beliefs about AI is the right starting point for value setting. This would probably increase pluralism and legitimacy in AI policies. The challenge is to target the appropriate category of evaluative beliefs. Given their individualistic nature, personal preferences revealed in natural environments might be irrational, unreliable, and hard to aggregate into collective preferences (see [24]). Rather than personal preferences, we argue that AI goals should be grounded in a different class of evaluative beliefs: moral intuitions.

Even though scholars in AI ethics (e.g., [51]) stress the relevance of moral intuition to inform AI, no one provides a specific psychological characterization of this mental state. Following recent research in moral psychology [9], we understand intuition as a specific type of moral judgment that possesses the following features. First, moral intuitions are defined by their *moral content*, a certain proposition asserting that something or someone is right, wrong, good, bad, morally obligatory, or permissible. The level of generality of the moral content is various: people can have intuitions about particular cases (e.g., that torturing a cat for fun is wrong), general judgments (e.g., that *ChatGPT* ought to disallow prompts about constructing lethal weapons), mid-level principles (e.g., that the development of AI should promote justice and minimize all types of discrimination), or abstract theoretical principles (e.g., that the rightness of an action depends on its consequences). Second, moral intuitions are *automatic* moral responses because they derive from largely autonomous processes—that is, involuntary, fast, and effortless [4, 19]. Automaticity captures the spontaneous and immediate aspect of moral intuition, distinguishing it from slower and effortful reflective judgments. Third, moral intuitions are also *strong* mental states insofar as they are experienced with a substantial degree of confidence, as compared to "shallow" automatic responses, such as guesses or quick hypotheses [6, 9]. Importantly, the strength of moral intuitions inclines the subjects to assent to their content and motivates them to act accordingly.

AI ethicists require scientific tools to collect intuitions with these specific psychological features. Although correlational studies and online surveys (such as [2]) offer some insight into people's preferences about AI, only well-designed psychological experiments can gather stable resposes by controlling the environment and excluding confounds [41]. Granted, we acknowledge the potential of a plurality of experimental conditions, such as qualitative interviews, self-report questionnaires, and observational studies.

In line with our hybrid approach, we contend that researchers should use experimental tools to collect various types of moral intuitions to the extent that each plays a different role in AI alignment. General principles like "The development of AI should ultimately promote the well-being of all sentient creatures" are important for defining ethical guidelines and policies. In contrast, intuitions based on specific situations accomplish the function of exemplifying, challenging, or testing general statements.[8] Abstract theoretical principles do not directly inform ethical AI but do play a role in justifying ethical principles and policies. For example, the statement that AI should promote well-being finds support in the utilitarian principle that an action ought to maximize general welfare.

Each type of intuition has strengths and limitations. Abstract intuitions tend to find more consensus across different cultures and are useful for integrating particular moral judgments into general goals for AI. However, they tend to be vague and there is the risk of overgeneralization from typical cases.[9] Particular intuitions, by contrast, are more subject to disagreement but are important in the application of principles.[10] For these reasons, a comprehensive value setting for AI would consider intuitions of all types and no priority should be given to a certain intuition only for the level of generality of its content.[11]

One might object that intuitions are not a promising starting point for AI alignment because substantial evidence shows that moral judgments are subject to social, gender, personal, and cognitive biases (see [34], pp. 45–89 for a review). If moral intuitions are biased, the objection goes, they are no more conducive to beneficial AI than personal preferences. However, we contend that their intrinsic psychological features and the methods by which they are collected make moral intuitions a suitable target for AI alignment.

One important aspect to consider is that intuitions, unlike personal preferences, have *moral content*. There are good reasons to believe that encouraging people to adopt a moral point of view fosters agreement on social problems related to AI. This point presupposes that morality binds individuals together rather than exacerbating value conflicts. Recent developments in moral psychology support this hypothesis, drawing on convergent evidence from evolutionary psychology and cultural anthropology [12, 53]. This evidence suggests that moral judgment inclines individuals toward solutions to cooperation-related problems inherent in human social life. Specifically, this line of research emphasizes that cooperative behaviors such as aiding one's group, reciprocating costs and benefits, or fairly distributing resources tend to be universally regarded as morally good across diverse cultures and ethical systems, while uncooperative behavior is universally considered morally undesirable. Therefore, if the theory of morality as cooperation is correct, as current evidence suggests, subjects induced to judge morally will be inclined to find cooperative solutions to problems regarding AI, accommodating multiple individual needs rather than satisfying personal desires.[12] In brief, enhancing ethical judgments can favor a shift from a competitive to a *cooperative* conception of AI. Note that this is compatible with ethical disagreement about how to understand cooperation (see Curry et al. [13]) and we will come back to this question in the next section.

To elicit cooperative moral attitudes, researchers can rely on the relevant psychological mechanisms underlying moral intuitions, such as moral emotions like sympathy or a sense of justice [26]. Importantly, the automatic and spontaneous nature of intuition fosters the participation of non-expert subjects in value setting. To elicit a particular intuition no sophisticated ethical knowledge is required, but presenting a morally salient case is sufficient. However, the automaticity of emotion and intuition does not exclude responsiveness to reasons [36, 50]. Rather, some empirical evidence suggests that certain moral principles (e.g., the doctrine of double effect) can be operative in non-expert moral intuitions, although the subjects fail to articulate them afterward [27]. In line with these findings, we assume here that moral intuitions can be sensitive to reasons even if not followed by accurate post-hoc justifications. Therefore, targeting moral intuitions for the AI alignment process has the potential to include a wide range of evaluative viewpoints in society. Researchers can use analytical tools to compare intuitions collected from different social groups (e.g., students, workers, philosophers, AI experts, etc.) and geographic areas to identify cross-cultural data.

Besides moral content and automaticity, the strength of moral intuitions is also relevant for AI alignment. Specifically, substantial empirical evidence shows that strong confident moral judgments (i.e., intuitions, according to our definition) tend to be stable over time and across circumstances

[8] General intuitions might be tested by qualitative methods that elicit reflection on ethical issues in AI (e.g., [16] and [40]). Instead, particular intuitions may require quantitative measurements of moral judgment in response to specific scenarios involving AI (e.g., [20]).

[9] In support of these statements, see the already mentioned review by Jobin et al. [30]. For the claim that general intuitions tend to be more stable, see Dabbagh [14].

[10] For example, in the ethics of autonomous vehicles, the principle of the *Institute of Electrical and Electronics Engineers* "to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression" [29] has been challenged the particular intuition to prioritize the young over the elders when presented an avoidable accident [2, 20].

[11] We disagree here with Huemer [28], according to which general moral intuitions are less prone to biases.

[12] A recent study investigating algorithmic interpretability and transparency corroborates this hypothesis [59]. In the study, participants are asked to justify the implementation of an algorithm to allocate limited resources in different real-life scenarios; although the subjects opt for different solutions, moral concepts like "fairness" or "rightness" mostly guided their decisions.

(see [9] for a review). In other words, subjects are less likely to revise strong intuitions than shallow evaluative preferences. Accordingly, moral intuitions tend to truly represent people's core moral values, and this constitutes an apparent reason to consider intuitions in an alignment process that endeavours to make AI universally beneficial. To distinguish intuitions from mere guessing, researchers can rely on experimental data such as self-reported "feeling of rightness" [56] or emotional arousal to measure subjects' confidence about their moral judgment.

Although they represent significant steps forward, moral content, automaticity, and intuitive strength are insufficient to minimize human biases. One might still object that, even if a methodological procedure provides an accurate idea of society's moral view, this view could still be fallacious, that is, completely off the track from beneficial AI. The quality of some output judgments, the objection goes, ultimately depends upon the quality of the inputs. In short: *garbage in, garbage out*. Though moral intuitions can be sensitive to reasons, this does not mean that every intuition is rational and reasonable. Even if accurately collected, intuitions can still be racist or misinformed and, hence, not conducive to beneficial AI. In response to this potential criticism, we emphasize the existence of epistemic tools that can improve the quality of intuitions by mitigating certain biases. Specifically, we refer to the most advanced *debiasing strategies* already in use in cognitive and social sciences. The goal of these tools is to provide optimal conditions to judge moral problems. We provide here some examples.

Cognitive biases like framing effects, overconfidence, or hindsight bias can be significantly reduced by intervening in information salience and presentation of a moral problem. Presenting a situation clearly and fairly and ensuring that the subjects understand the relevant information can improve the quality of moral intuitions. For instance, some empirical evidence suggests that simply drawing subjects' attention to the importance of framing a problem reduces framing effect, without involving subjects' analytical reasoning [5]. Overconfidence, instead, can be mitigated by informing subjects that a moral problem is debated and controversial [61] or inviting them to consider opposite solutions [35]. Finally, hindsight bias can be effectively reduced by presenting to subjects alternative outcomes of a probabilistic event or adding to the problem presentation an expert's opinion about the actual probability of a harmful outcome [32].

Social biases can also be mitigated by appropriate experimental design. Indeed, long-standing research in social psychology shows that implicit prejudices are remarkably malleable according to the local environment in which a subject is immersed [15, 23]. Thus, similarly to cognitive debiasing, researchers can generate an artificial environment that minimizes social biases and temporarily frees participants from their influence. For this purpose, for example, experimenters can favor participants' awareness of possible sources of prejudice or using goals and motivations to weaken implicit stereotypic associations [23]. Furthermore, a large body of evidence suggests that exposing subjects to counterstereotypic environmental cues (e.g., displaying images of admired and famous African Americans) strongly reduces implicit biases [15].

In summary, collecting moral intuitions is a promising starting point for AI value setting. Compared to personal preferences, eliciting people's ethical intuitions would foster a cooperative rather than a competitive understanding of AI. Although moral intuitions are not extremely reliable per se (especially in everyday environments), we are moderately optimistic that suitable experimental settings can increase their reliability by debiasing strategies.

4.2 Justifying moral intuitions and reasonable intuition conflict

Once collected and filtered by suitable methods, moral intuitions need to be explained and justified in order to infer meaningful ethical demands for AI. We define *intuition justification* as the reflective process of articulating normative reasons for the content of some relevant intuitions. This practice encompasses activities such as connecting specific moral judgments with general principles or constructing arguments for ethical AI based on evidence and relevant background theories. Importantly, the ultimate goal of this process is to transform implicit moral information contained in intuitions into an evidence-based cooperative understanding of AI.

The process of articulating and exchanging normative reasons brings moral intuitions into an intersubjective dimension in which researchers communicate ethical views on AI to the scientific community. According to recent evidence, this aspect tends to enhance the quality of output intuitions, mitigating certain individual cognitive biases. The reasons are manifold [37]. Firstly, receiving feedback from a scientific audience is beneficial because individuals tend to be more accurate in evaluating others' arguments than their own ([37, p. 231]). This helps reduce the documented "myside bias," which refers to the systematic difficulty individuals face in seeking counterevidence and counterarguments to their own opinions. Secondly, communicating intuitions is beneficial for reasoning insofar as, during interactive discussions, individuals exchange numerous concise arguments. This facilitates the development of longer and more robust arguments with minimal individual cognitive effort ([37, p. 224]). Thirdly, and finally, intuition communication tends to promote coherent justifications as it is easier to persuade an interlocutor of a claim by demonstrating its coherence with their existing beliefs ([37, p. 194]).

The task of intuition justification has traditionally fallen within the purview of normative theory and there appears to be no reason why it should not continue to do so in the ethics of AI. We also emphasize the complementary role that recent moral-psychological theories can play in illuminating the underlying principles behind moral intuitions. For example, according to *Moral Foundation Theory*, moral intuitions regarding AI diverge based on the emphasis individuals place on different fundamental moral concepts ("moral foundations") [54].[13] In contrast, the *Agent, Deed, and Consequences* (ADC) model of moral judgment proposes that intuitions about AI vary according to the character and intentions displayed by an artificial agent (the Agent component, *A*), the intrinsic nature of its actions (the Deed component, *D*), and the outcomes produced (the Consequences component, *C*) [44].[14] Relevant to our purposes, moral-psychological theories can contribute to understanding intuition conflict and integrating competing normative standpoints.

While the combination of normative and moral-psychological theory can mitigate biases and resolve intuition conflicts, we are not sufficiently optimistic to claim that every ethical conflict can be solved solely by moral reasoning.[15] Instead, it is more realistic to expect that some reasonable disagreement regarding AI will persist even after intuition justification. This is likely due to the irreducible complexity of moral problems and the existence of multiple conceptions of cooperative AI. As previously mentioned, though people tend to universally agree that "ethical AI" entails "cooperative AI", there exist varying interpretations of cooperation and diverse cooperative approaches to addressing social issues related to AI. These differences give rise to competing ethical perspectives. In what follows, we illustrate some paradigmatic examples.

An AI application that pervades our everyday lives is marketing algorithms. Moral intuitions about their deployment tend to be polarized between two opposing views [3]. A free market-oriented view emphasizes the freedom of stakeholders to make voluntary agreements and earn economic benefits; accordingly, consumers should be relatively free to share personal information in exchange for gain, while informed consent and "opt-out" policies should suffice to regulate agreements. By contrast, intuitions more focused on consumer protection highlight the risks to privacy and individual autonomy, advocating for restrictive regulations (e.g., "opt-in" policies) to mitigate the sharing of personal data. Another paradigmatic case of intuition conflict can be identified in the ethics of driving automation. In this field, intuitions are divided between a public transportation-centered vision and a private ownership-centered vision [16]. Whereas the former view underlines the potential of automated public vehicles to reduce environmental impact and ensure citizens' right to independent movement regardless of income, the private ownership vision highlights the importance of private cars for freedom of mobility.

While a systematic investigation of ethical disagreements in AI applications falls outside the scope of this paper, the examples described above suffice to illustrate what we define as *reasonable intuition conflict*. Given two intuitions I_1 and I_2, a reasonable conflict between them occurs whenever:

- I_1 and I_2 favor opposing cooperative views of an AI application
- Both I_1 and I_2 are provably unbiased (i.e., not resulting from mere personal interests, social prejudice, or systematic mistakes in interpreting information)
- Both I_1 and I_2 are supported by compelling normative reasons.

In other words, reasonable intuition conflict is an ethical disagreement that persists despite the mitigation of biases and the justification process. Finding trade-offs between reasonable intuition conflicts is a task primarily for politics, rather than empirically informed ethical inquiry, or so we argue. The optimal outcome we expect from the normative justification of intuitions consists of a limited set of refined cooperative models of AI applications, each stemming from reliable moral intuitions. This class of intuitions would serve as a starting point for public discussion and political deliberation aimed at incorporating intuitions into AI policies.

Detailing how public discussion and politics should embody moral intuitions about AI goes beyond the purposes of this paper. It is plausible that trade-offs between intuition conflicts should be sensitive to the cultural and political context of the geographical area where an AI application is deployed. For instance, the political history of the United States suggests that this region may be more receptive to a free market view of marketing algorithms and a private ownership-centered vision of driving automation. By contrast, the European Union, historically more inclined toward privacy protection and a welfare system, may lean toward a consumer protection view of marketing algorithms and a public transportation-centered vision of driving automation.

[13] For instance, a person inclined to the "authority" foundation might approve extensive data collection for security purposes, while a subject sensitive to "liberty" could not see that as a sufficient reason for the privacy intrusion [54].

[14] For example, moral intuitions about a self-driving car's decisions vary according to the weight given to the car's driving style and reliability (A), the compliance with traffic norms (D), and whether the action results in an accident (C) [10].

[15] We agree with Savulescu et al. [51] on the fact that "overlapping consensus" between intuitions from different sources is desirable and should be strongly considered in AI policy making. However, we assume here that consensus is not always possible and our discussion focuses on cases of reasonable disagreement.

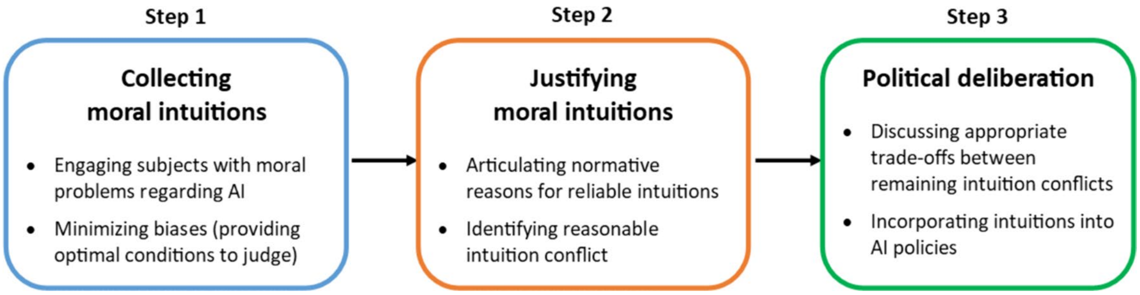
Step 1
Collecting moral intuitions
Engaging subjects with moral problems regarding AI
Minimizing biases (providing optimal conditions to judge)
Step 2
Justifying moral intuitions
Articulating normative reasons for reliable intuitions
Identifying reasonable intuition conflict
Step 3
Political deliberation
Discussing appropriate trade-offs between remaining intuition conflicts
Incorporating intuitions into AI policies

Fig. 1 The intuitionist approach in three steps

5 Discussion

Our goal was to detail a hybrid methodological procedure conducive to beneficial AI. Before considering the advantages and limitations of our intuitionist account, let us summarize the main highlights. In brief, we contend that AI systems should embody potential stakeholders' moral intuitions. Thus, AI value setting should follow three main steps (Fig. 1). First, researchers should collect reliable moral intuitions via appropriate experimental designs. The goal of this stage is not to neutrally register people's moral views but to take "the best from them," providing optimal conditions to judge thanks to effective debiasing strategies. The second step consists of the justification of intuitions by normative and psychological theories. Crucial for this stage is the articulation of normative reasons and the identification of reasonable conflicts. Even this stage should enhance the quality of intuitions by spelling out their rationale and transforming their implicit content into cooperative views of AI. Finally, the third step of our research agenda comprises the political deliberation of values for AI. The purpose of this stage is to incorporate moral intuitions into AI policies and discuss appropriate trade-offs between the remaining intuition conflicts.

The account outlined above combines the strength of bottom–up and top–down approaches while avoiding the issues that emerged in Sect. 3. Unlike top–down methods, the intuitionist approach enhances value pluralism by incorporating in the procedure a wide spectrum of moral opinions, including experts and non-experts from a variety of social, racial, and geographic demographics to promote optimal representation. Yet, the intuitionist method also aligns with top–down methodologies as it concerns the ethical cooperative approach and the use of expertise to filter and justify individual beliefs. These aspects position our method to better integrate individual values into universal goals. Specifically, unlike bottom–up approaches, we have identified better input (moral intuition instead of personal preferences) and more reliable strategies to enhance the quality of the inputs (experimental designs and moral theory instead of machine learning). Therefore, our proposed procedure can reconcile inclusiveness and output reliability in AI alignment.

The intuitionist approach complements existing hybrid methodologies [51, 57] by indicating the kind of individual beliefs apt for informing AI and the psychological tools to collect and filter them. We have also introduced the novel concept of reasonable intuition conflict and provided necessary and sufficient conditions for it. Nevertheless, we acknowledge that our account is less ambitious than VSD [57], not covering the full cycle of AI alignment (encompassing value setting, implementation, and feedback).

Our conclusion may appear skeptical to the extent that we do not expect that moral theorizing can solve every kind of conflict between intuitions. However, we do believe that minimizing biases and understanding reasonable disagreements represent a realistic yet important achievement for empirically informed ethics. We are cautiously optimistic that distinguishing apparent (due to biases) and reasonable intuition conflict would put politics in a favorable position to find legitimate trade-offs between values. We do not address at length this political aspect of AI alignment, at the interface between value setting and value implementation. Discussing how AI policies should incorporate moral intuitions could be a subject for future work.

The intuitionist approach largely relies on experimental settings to enhance the quality of moral opinions. We anticipate here some concerns regarding the recognized limitations of experimental ethics [33]. In particular, two pressing issues have emerged in the recent literature: firstly, moral studies' lack of ecological validity, which pertains to the potential discrepancy between experimental settings and real-world moral situations. Consequently, moral intuitions collected under such conditions might not accurately reflect people's core values. Secondly, like other experimental disciplines, empirical ethics grapples with a *replicability crisis*, meaning that many empirical results fail to be replicated by independent researchers.

As concerns the problem of ecological validity, we stress how some technological resources can enhance the realism of experimental moral scenarios, thereby increasing the psychological engagement of participants. For example, a relatively recent tool that can tremendously increase a moral study's ecological validity is virtual reality (VR). This technology permits AI ethicists to use naturalistic immersive depictions of moral situations to test moral judgment. VR studies can be particularly beneficial for informing AI because experimenters can observe how people interact with AI agents in a realistic environment that can be controlled and manipulated to obtain accurate data. Specifically, VR enables various useful conditions to investigate moral judgment: external observers' judgment, simulated decision-making, and interactions between multiple agents. For such reasons, a growing number of studies have applied VR to road traffic scenarios to contribute to the ethics of automated vehicles [10, 20]. Nevertheless, VR environments can be adapted to study moral intuitions in other domains of AI.

The replicability crisis can be tackled by adopting good research practices that are increasingly becoming standard in empirical disciplines. For instance, increasing transparency and clarity in reporting methods and results would facilitate future replications. Furthermore, it is essential not to solely rely on single statistical parameters (e.g., the *p*-value) and to report complementary measures [45]. Finally, in the ethical domain particularly, it is crucial to not overgeneralize from mixed results, as they may indicate reasonable normative disagreement [48].

Another potential objection to our research agenda concerns the assumption of some epistemic and moral norms in bias mitigation. For example, social prejudice reduction is justified by respecting human dignity (thus, no discrimination). On the other hand, reducing cognitive biases presupposes adherence to basic standards of rationality embedded in scientific practice. While some may view these assumptions as risking manipulation of subjects' opinions or as employing a circular methodology (see [47]), we contend that assuming some fundamental normative truths is inevitable to situate AI alignment within the framework of liberal democracies and scientific practice, which still leave room for reasonable disagreement about more specific values. Furthermore, we emphasize that achieving complete value neutrality in collecting intuitions is hardly feasible. Even in presenting certain moral problems to subjects, researchers must select information considered relevant, a process that inherently carries value-laden implications. Hence, it might be preferable to adopt norms that are consolidated in liberal democracy and scientific practice rather than implicitly accepting arbitrary values. Admittedly, what is bias is debatable to some extent and, therefore, we encourage further discussion in future work.

This paper has primarily focused on AI value setting while setting aside the issue of value implementation, which encompasses integrating selected values into AI systems. At these preliminary stages, the challenge we foresee in encoding moral intuitions lies in devising mechanisms within artificial agents that can replicate the psychological processes underlying moral intuitions, such as moral emotions or mental heuristics. Replacing such mental processes with functional equivalents would facilitate the design of AI systems capable of making decisions aligned with human moral intuition.

6 Conclusion

The explosion of highly autonomous artificial agents raises the ethical challenge of aligning their goals with society's values. Given the existence of value pluralism, a reliable procedure is demanded to determine the values with which AI ought to align. The goal of such a methodology is to include a wide range of existing evaluative viewpoints in society and, at the same time, select and integrate them to make future AI beneficial.

In light of the strengths and limitations of existing top–down and bottom–up approaches, we have described a hybrid intuitionist approach to the alignment problem. We have proposed a research agenda that is inclusive enough to consider moral intuitions from all potential AI stakeholders while ensuring the reliability of the output values. Finally, we have discussed the process of intuition justification and the conditions for reasonable intuition conflict.

Focusing on the methodology for empirical ethical inquiry, this paper left mostly unaddressed further phases of AI alignment such as the incorporation of moral intuitions in policymaking and how they can be technically embedded into AI algorithms. These are tasks for future work.

Author contributions D.C.: ideation, conceptualization, and draft preparation. M.P.: review and editing. V.D.: supervision, review, and editing.

Funding Funding was provided by NSF Division of Social and Economic Sciences (grant no. 2043612, awarded to V.D).

Declarations

Conflict of interest No conflict of interests.

References

1. Anderson, M., Anderson, S.L.: Case-supported principle-based behavior paradigm. In: Trappl, R. (ed.) A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations, pp. 155–168. Springer, Cham (2015)
2. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., ShariffRahwan, A.J.-F., Bonnefon, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)

3. Baase, S., Henry, T.M.: A Gift of Fire: Social, Legal, and Ethical Issues for Computing Technology. Pearson, New York (2018)
4. Bargh, J.A.: The ecology of automaticity: toward establishing the conditions needed to produce automatic processing effects. Am. J. Psychol. **105**(2), 181–199 (1992)
5. Baumer, E.P.S., Polletta, F., Pierski, N., Gay, G.K.: A simple intervention to reduce framing effects in perceptions of global climate change. Environ. Commun. **11**(3), 289–310 (2017)
6. Bengson, J.: The intellectual given. Mind **124**(495), 707–760 (2015)
7. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The moral psychology of ai and the ethical opt-out problem. In: Liao, S.M. (ed.) Ethics of Artificial Intelligence, pp. 109–126. Oxford University Press, Oxford (2020)
8. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6397), 36–37 (2016)
9. Cecchini, D.: Moral intuition, strength, and metacognition. Philos. Psychol. **36**(1), 4–28 (2023)
10. Cecchini, D., Brantley, S., Dubljević, D.: Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles. AI Soci (2023). https://doi.org/10.1007/s00146-023-01813-y
11. Christian, B.: The Alignment Problem: Machine Learning and Human Values. W.W. Norton & Company, New York (2020)
12. Curry, O.S., Mullins, D.A., Whitehouse, H.: Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. Curr. Anthropol. **60**(1), 47–69 (2019)
13. Curry, O.S., Alfano, M., Brandt, M.J., Pelican, C.: Moral molecules: morality as a combinatorial system. Rev. Philos. Psychol. **13**, 1039–1058 (2021)
14. Dabbagh, H.: Intuitions about moral relevance—good news for moral intuitionism. Philos. Psychol. **34**(7), 1047–1072 (2021)
15. Dasgupta, N.: Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. In: Devine, P., Plant, A. (eds.) Advances in Experimental Social Psychology, vol. 47, pp. 233–279. Academic Press, Burlington (2013)
16. Dubljević, V., List, G., Milojevich, J., Ajmeri, N., Bauer, W.A., Singh, M.P., Bardaka, E., et al.: Toward a rational and ethical sociotechnical system of autonomous vehicles: A novel application of multi-criteria decision analysis. PLoS ONE **16**(8), e0256224 (2021)
17. Dung, L.: Current cases of AI misalignment and their implications for future risks. Synthese **202**, 138 (2023)
18. European Union: Artificialintelligenceact.eu. https://artificialintelligenceact.eu/. Accessed May 2024 (2024)
19. Evans, J., Stanovich, K.: Dual-process theories of higher cognition: advancing the debate. Perspect. Psychol. Sci. **8**(3), 223–241 (2013)
20. Faulhaber, A.K., Dittmer, A., Blind, F., Wächter, M.A., Timm, S., Sütfeld, L.R., Stephan, A., Pipa, G.: Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci. Eng. Ethics **25**, 399–418 (2019)
21. Floridi, L.: The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, Oxford (2023)
22. Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018)
23. Forscher, P.S., Lai, C.K., Axt, J.R., Ebersole, C.R., Herman, M., Devine, P.G.: A meta-analysis of procedures to change implicit measures. J. Personal. Soc. Psychol. Attitudes Soc. Cognit. **117**(3), 522–559 (2019)
24. Gabriel, I.: Artificial Intelligence, values, and alignment. Mind. Mach. **30**, 411–437 (2020)
25. Hager, G.D., Drobnis, A., Fang, F., Ghani, R., Greenwald, A., Lyons, T., Parkes, D.C., Schultz, J., Saria, S., Smith. S.F.: Artificial intelligence for social good. arXiv:1901.05406 (2019)
26. Haidt, J.: The Moral Emotions. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (eds.) Handbook of Affective Sciences, pp. 852–870. Oxford University Press, Oxford (2003)
27. Hauser, M., Cushman, F., Young, L., Jin, K., Mikhail, J.: A dissociation between moral judgments and justifications. Mind Lang. **22**(1), 1–21 (2007)
28. Huemer, M.: Revisionary Intuitionism. Soc. Philos. Policy **25**(1), 368–392 (2007)
29. IEEE: IEEE code of ethics. https://www.ieee.org/about/corporate/governance/p7-8.html. Accessed Jun 2023 (2020)
30. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019)
31. Jonker, J.D.: Automation, alignment, and the cooperative interface. J. Ethics 1–22 (2023). https://doi.org/10.1007/s10892-023-09449-2
32. Kneer, M., Skoczen, I.: Outcome effects, moral luck and the hindsight bias. Cognition **232**, 1–21 (2023)
33. Luetge, C., Rusch, H., Uhl, M.: Experimental Ethics: Toward an Empirical Moral Philosophy. Palgrave Macmillan, Houndmills, Basingstoke (2014)
34. Machery, E.: Philosophy Within Its Proper Bounds. Oxford University Press, Oxford (2017)
35. Mata, A.: Social metacognition in moral judgment: decisional conflict promotes perspective taking. J. Pers. Soc. Psychol. **117**(6), 1061–1082 (2019)
36. May, J.: Regard for Reason in the Moral Mind. Oxford University Press, Oxford (2018)
37. Mercier, H., Sperber, D.: The Enigma of Reason. Harvard University Press, Cambridge (2017)
38. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nature Machine Intelligence **1**, 501–507 (2019)
39. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Moekander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI ethics. Minds Mach. **31**, 239–256 (2021)
40. Dubljević, V., Douglas, S., Milojevich, J., Ajmeri, N.: Moral and social ramifications of autonomous vehicles: a qualitative study of the perceptions of professional drivers. Behav. Inf. Technol. **42**, 1271–1278 (2023). https://doi.org/10.1080/0144929X.2022.2070078
41. Morling, B.: Research Methods in Psychology: Evaluating a world of information. Norton & Company, New York (2018)
42. O'Neil, C.: Weapons of Math Destruction. Crown, New York (2016)
43. OpenAI: GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed May 2023 (2023)
44. Pflanzer, M., Traylor, Z., Lyons, J.B., Dubljevic, V., Nam, C.S.: Ethics in human-AI teaming: principles and perspectives. AI Ethics **3**, 917–935 (2022)
45. Polonioli, A., Vega-Mendoza, M., Blankinship, B., Carmel, D.: Reporting in experimental philosophy: current standards and recommendations for future practice. Rev. Philos. Psychol. **12**, 49–73 (2021)
46. Rahwan, I.: Society-in-the-loop: programming the algorithmic social contract. Ethics Inf. Technol. **20**, 5–14 (2018)
47. Rini, R.: Debunking debunking: a regress challenge for psychological threats to moral judgment. Philos. Stud. **173**, 675–697 (2016)
48. Rosenthal, J.: Experimental philosophy is useful—but not in a specific way. In: Luetge, L., Rusch, H., Uhl, M. (eds.) Experimental Ethics: Towards an Empirical Moral Philosophy, pp. 211–226. Palgrave Macmillan, Houndsmill, Basingstoke, Hampshire (2014)
49. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin, New York (2019)

50. Sauer, H.: Moral Judgments as Educated Intuitions. MIT Press, Cambridge (2017)
51. Savulescu, J., Gyngell, C., Kahane, G.: Collective reflective equilibrium in practice (CREP) and controversial novel technologies. Bioethics **35**, 652–663 (2021)
52. Seligman, M.E.P.: Flourish : A Visionary New Understanding of Happiness and Well-Being. Free Press, New York (2011)
53. Sterelny, K., Fraser, B.: Evolution and moral realism. British Journal of the Philosophy of Science **68**(4), 981–1006 (2016)
54. Telkamp, J.B., Anderson, M.H.: The implications of diverse human moral foundations for assessing the ethicality of artificial intelligence. J. Bus. Ethics **178**, 961–976 (2022)
55. The White House: Blueprint for an AI bill of rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed May 2024 (2022)
56. Thompson, V., Turner, J.P., Pennycook, G.: Intuition, reason and metacognition. Cogn. Psychol. **63**, 107–140 (2011)
57. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics **1**, 283–296 (2021)
58. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
59. Webb, H., Patel, M., Rovatsos, M., Davoust, A., Ceppi, S., Koene, A., Dowthwaite, L., Portillo, V.: "It would be pretty immoral to choose a random algorithm": opening up algorithmic interpretability and trasparency. J. Inf. Commun. Ethics Soc. **17**(2), 210–228 (2019)
60. Wong, D.: Moral Relativity. University of California Press, Berkeley (1984)
61. Wright, J.C.: Tracking instability in our philosophical judgments: is it intuitive? Philos. Psychol. **26**(4), 485–501 (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.