How to gain control and influence algorithms: contesting AI to find relevant reasons

Kuilman, S.K.; Cavalcante Siebert, L.; Buijsman, S.N.R.; Jonker, C.M.

DOI
10.1007/s43681-024-00500-y

Publication date
2024

Document Version
Final published version

Published in
AI and Ethics

AI and Ethics
https://doi.org/10.1007/s43681-024-00500-y

ORIGINAL RESEARCH

How to gain control and influence algorithms: contesting AI to find relevant reasons

Sietze Kai Kuilman[1] · Luciano Cavalcante Siebert[1] · Stefan Buijsman[1] · Catholijn M. Jonker[1]

Received: 29 March 2024 / Accepted: 15 May 2024

Abstract
Relevancy is a prevalent term in value alignment. We either need to keep track of the relevant moral reasons, we need to embed the relevant values, or we need to learn from the relevant behaviour. What relevancy entails in particular cases, however, is often ill-defined. The reasons for this are obvious, it is hard to define relevancy in a way that is both general and concrete enough to give direction towards a specific implementation. In this paper, we describe the inherent difficulty that comes along with defining what is relevant to a particular situation. Simply due to design and the way an AI system functions, we need to state or learn particular goals and circumstances under which that goal is completed. However, because of both the changing nature of the world and the varied wielders and users of such implements, misalignment occurs, especially after a longer amount of time. We propose a way to counteract this by putting contestability front and centre throughout the lifecycle of an AI system, as it can provide insight into what is actually relevant at a particular instance. This allows designers to update the applications in such a manner that they can account for oversight during design.

Keywords Relevancy · Value alignment · Frame problem · Contestability

1 Introduction

Artificial intelligence (AI) systems should be aligned towards societal good. Nonetheless, mistakes are racking up, and so there are attempts to get a better grip on how to implement and control AI systems. The result of this search for better AI systems and control over them is not without merit, as it has led to numerous theories about how one ought to control and/or design such systems [17, 19, 36, 43]. There are also good reasons to search for a solution to misalignment, as such systems cause serious harm [5, 12, 42] or prove detrimental to institutions [8].

Much of the troubles designers have with their implementation has to do with moral reasons. Designers need to weigh or trade-off certain values, understand a treasure trove of contextual information, and discern between a set of minute details that may not at all be clear at first glance. The saying: the devil is in the details, could not be more true with regard to value-alignment. At the end of the day it is not the high and mighty ideals of the designer that matter, but the down on the ground implementation affecting people's lives, and when there is serious harm, then these may also inadvertently skew the public debate such that AI systems can become unwanted.

What we aim for, in this paper, is to alleviate some of the issues involved with implementing value aligned strategies. As we shall show, in practice it may be difficult to determine what reasons are relevant and which are not. This is not only due to contextual factors, but also who is involved in the design process. We propose to include contestation at different stages of the system design to improve the situation and application of AI systems, and we offer a mostly agnostic way to adjust the system based on missing inferences.

In short, we will first go over some difficulties with the implementation of value alignment. What we will see is that it requires knowing relevant moral reasons. Yet, these

✉ Sietze Kai Kuilman
S.k.kuilman@tudelft.nl

Luciano Cavalcante Siebert
L.CavalcanteSiebert@tudelft.nl

Stefan Buijsman
S.N.R.Buijsman@tudelft.nl

Catholijn M. Jonker
C.M.Jonker@tudelft.nl

[1] Delft University of Technology, Delft, The Netherlands

Published online: 05 June 2024

are difficult to find because relevancy is not a given. If we assume that to be the case, then we implicitly make assumptions about what is relevant and what is not, likely leading to unaligned implementation. The issues with relevancy are two-fold: (1) Theoretically, formalization has certain issues and (2) Designers are limited in knowing what is relevant, even if they talk to stakeholders. To counteract this, we introduce contestation throughout an AI lifecycle to better align such systems, but this requires that contestation leads to meaningful adaptation.

2 When we talk about relevancy

Value alignment may be difficult to implement because of relevancy. We need to find the relevant moral reasons, but if a designer does not know those or understand them and has no means of finding them, they will be hard-pressed to ever create a value aligned system. Relevancy is, in that regard, the bedrock of most value aligned theories, as they deal with the matter of correctness. It is not always transparent what and when something is relevant, so without any good substantiation of such a notion we can provide beautiful theories with little to no effective application, as we may not have the means to find the correct moral reasons for a given situation.

Of course, we need to first understand what we mean by moral reasons. The nature of a moral reason is dependent on one's views of morality, but to keep a general statement about moral reasons. One can view them as a determinant which plays a part in one's action. For example, Lying is wrong, gives us a duty not to lie, but the reason why we ought not lie could be: never treat another as a means to an end. This is of course a very different presupposition compared to: honesty maximizes happiness. Both leading to the behaviour of honesty, but in particular situations these underlying beliefs matter for how a problem is approached and what particulars are used in a solution. Moral reasons are thus about the domain of correct action (e.g. prevention of harm). Normativity is a broader in scope, which could be addressed here as well, but we dive in particular into the notion of what one ought to do.

A central topic of moral reasoning is also casuistry, the search of finding the relevant moral conditions and discerning the more relevant ones from the only slightly relevant ones. If a designer is incapable of making relevant distinctions in context or understand what needs to be kept track of in said context, then it doesn't make a difference if they try. Furthermore, if those relevant distinctions are not presentable in such a system (e.g. those features cannot be captured because of their complexity or inherent indescribability), it will most likely cause the same problem. The question for correct behaviour is thus:

What relevant moral reasons do we require for such AI systems to function properly?

For computer science, relevancy has been of much importance within information retrieval, as finding the correct and relevant document has been vital to the field [37]. If the relevancy of a document is determined by the amount of clicks it gathers (say on the website of a search engine), then we can safely say that older documents which have been exposed to time are likely to have generated more clicks, in that regard an underlying belief is that older documents may be more relevant. Certainly we can understand this for literature, where classics have to withstand the test of time, but this may not be appropriate for certain types of documents (say scientific information). The quest for a designer is to know which reasons are relevant and how they should be implemented such that the right documents fall into the right hands.

In this paper, we will take one route to relevancy, but we believe there are many more possibilities. However, the main premise we postulate in terms of relevancy is that during design we need to fill in the details. We make choices when applying a (moral) theory about what is relevant. We argue that, without an adequate idea and application of relevancy, we are at a loss of finding effective applications as the openness in implementation ends up detailing the most important part.

2.1 Value alignment

Theories on value alignment are built upon the proposition that AI should not merely act, but should also act such that certain harms can be avoided [18] or that they positively influence a kind of human flourishing [23, 33, 34]. How do we make sure that machines act in accordance with our values? There are two basic approaches to this: sociotechnical solutions [5, 16, 17, 36, 44]and more technical ones [14, 19, 22, 29]. Sociotechnical solutions involve users and try to picture the machinery such that there can be an interplay between humans and artificial intelligent agents. Technical solutions stem from the belief that value-aligned action can be seen as a technical problem and can be solved as such. While there are obvious differences between sociotechnical solutions and technical ones, the main premise of this paper holds for both approaches. The particulars of technical solutions, still requires data, features, and some line by which to draw what is relevant and what is not. The inference of a pattern from said data and features also requires much more understanding in concrete cases than simple assuming that this relevancy is easily found or induced appropriately.

Applying value alignment theories in any domain is a feat dependent on context, goal, and structure. For example, knowing whether an application is discriminatory requires that we also know whether that discrimination is at any point acceptable in that context, meaning that we need to disambiguate discrimination as discernment (to delineate different options) and discrimination as unjust bias (to categorize groups on features that are deemed inappropriate). We may want to discriminate (discern) between different groups, but we do not want to discriminate individuals (unjustly). This is essential, because we may sometimes really want to recognize a sick patient from a healthy one. However, we don't want a system to only recognize sick patients of a particular gender (unless the disease is gender-specific of course). Considering the context, we need to know whether we are actually introducing unjust bias or doing the right inference. This means we have to know at which juncture it is one or the other.

Such knowledge requires a particular kind of oversight and knowledge of the system. The problem of reward hacking in terms of goals [3]—that being, the AI system finds a misinterpretation of the goal such that it can maximize its reward function—presents a serious issue to this kind of knowledge. As an AI system may optimize for something (unintentionally) through unjust means. The context in which it is placed is also highly important. Facial recognition is not necessarily unaligned, but if it is applied in a way to arrest a particular group, then we can talk of the relevancy of being able to discern that group and the misalignment of that in the face of human flourishing.

On top of that, designers do make choices about what is relevant and what is not, both on a technical level and on a social one. If a designer wants a recommender system to be value aligned, then they need to know what ought to be recommending and what they can recommend. This entails a kind of idea about what the most relevant detail may be within possible documents and how to extract it. Such a detail needs to be discerned from the context, meaning that the context actually does need to contain that detail. If the data is structured in a way that does not allow correct or full access to the relevant detail—or does it in a way that ties it to other factors, then they are bound to infer a different pattern than the actual relevant one. A choice in recommending based on citations or recommending based on amount of views, is likely going to result in very different recommendations. While we can debate whether that is moral perse, this is easily transposed towards a moral domain by changing clicks and cites into recommendation based on sensitive topics. For example, demoting climate skepticism or certain information about war.

Not considering these issues is not a way out. To assume the data collected is correctly structured and always contains the correct scope of information, even humanly labelled data, to reach the correct goal is quite hefty assumptions to make. It does not mean value alignment becomes easier, in fact it only means that one has made implicit assumptions about what is relevant on a technical level and on a social level, without delving into it.

2.2 Relevancy

Having argued for the necessity of relevancy, we need to understand what we need to know about relevancy. However, this is somewhat difficult to define. When is a reason relevant?

Relevancy is found in a plethora of fields and studies, such as: communication [40], logics [4, 13], and information retrieval [26, 37]. It will most likely play a part in many goal-oriented studies. Finding the correct treatment means knowing certain relevant facts about a disease. Knowing how to construct policy means knowing (some of) the relevant actors in a particular case. To find the relevant moral reason requires knowing what they are or how to find and evaluate them. The point of constructing or excavating a concept like relevancy in value alignment is not a quest to survey all potential relevancies. Rather, we desire to know what kind of relevancy we could look to.

As we mentioned in the section on value alignment, there are contextual features, saliency, and teleology to keep in mind when discussing relevancy. How we use an AI system is of importance to value alignment. And while we could discuss the specifics and lack of oversight of applications as a serious issue, for this paper we will assume this is about intended use. Where it is used and for what purpose it is used, seem like obvious parts to relevancy. This however does not explicitly cover saliency. Yet, if a relevant distinction is not salient to the system, then it will likely not be able to draw the right inferences from the context.

AI systems hold a particular position within societal, governmental, or commercial institutions which is very different to that of humans and this needs to be kept in mind. As the relevancy of human beings may change and update on the fly, while that of a system (even online ones) have been trained in a particular fashion and have a certain depth and breath of possible implementations. During design, one needs to scope what the possible actions for an AI system: what its range of contextual features are and what is salient to it. In the broadest sense, in value alignment, we seem to be asking for the impossible. As Turing already noticed:

> It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality, even

> those arising from traffic lights, appears to be impossible. [41]

One should note that Turing's comment also works for learning relevancy in machine learning models. The decision of data and the model used within a set of circumstances may be inappropriate in another context. Such limitations make it impossible to produce the induction of a function (rather than a set of rules) which allows the machine to act in every conceivable set.[1]

In the ideal setting, we would be able to cover every eventuality, then our systems would be completely value aligned, acting appropriately in all edge cases and novel circumstances. Yet, we also need to see what relevancy means if we are pragmatic. To reiterate on the example of recommender systems from the previous section, if we desire that recommender systems do not spread misinformation, we need to evaluate what the content they recommend, and this shifts with time. The scientific community may gather a new view on things or a definition of misinformation may miss a particular new piece of misinformation due to its novelty. If we desire to value integrity and honesty (as values which we take to be at least some basis of why we want to avoid giving misinformation) then this boils down to defining what those values mean (not sharing misinformation). Yet, that is simply not a static conception nor a static output. Things that are misinformation may turn out worthwhile and vice versa. When something is relevant has thus much to do with its context and its intended use. A proper functioning of an AI system thus means that we can detail what exactly we need from it, in terms of delineating context and inference patterns in relation to its goal.

Yet as noted, relevancy of these systems comes in a particular way and this opens us up to two problems: the formalization aspect of such systems and those who are involved in the process of formalization. After we have delved into that topic in particular, we will spend the remainder of the paper explaining what we can do to find the relevant moral reasons.

3 Addressing the algorithm in the room

The main reasons why we distinguish between relevancy humans display and those that belong to artefacts result from the formalization of context and use, and the fact that these artefacts are being designed by designers. These factors lead to issues that make finding the relevant reasons far from easy. We will describe why this may be the case in this section.

3.1 The frame problem

In the sixties and seventies, a roughly similar problem as the one we describe with value alignment was addressed in terms of expert systems with logic statements. McCarthy and Hayes [25] recognized that there was a problem with representationalism, namely: there is a lack of inertia when dealing with predicates. Each time an update function was performed (to see if any predicates had changed based on action) all predicates had to be checked because there was simply no knowing which had to be updated and which had not. To think of this in simple terms, I can paint an object, but if I move it, how do I know it hasn't changed it colour? Going over each and every predicate was simply a waste of both calculations and space, because many things would not change given a simple action. Yet, not going over such dependencies might cause the machine to overlook simple yet important dependencies when predicates did change. They called this double-bind the frame problem.

The concept was quickly appropriated by a broader philosophical community, wherein the discussion was not specifically meant to address problems in logical calculus, but rather to address the question of relevancy and action [6, 10, 15]: How to act and update beliefs about the world? The "whole pudding" of the frame problem—meaning both the version McCarthy and Hayes defined and the ones philosophers aligned with it—shows the practical limits of describing actions in terms of relevancy. The frame problem applies today, even in machine-learning, as we can ask how we ought to define the world, and what limits we need to draw in featurizing such that the pattern we achieve is correct. This is not a trivial task, and it may be the reason why we resort to using a term like relevancy in value alignment. It may be too difficult to know which moral reasons we need to account for. To show this difficulty, we base ourselves on Daniel Dennett's example [10].

> One day its [R1] designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. ... There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called `PULLOUT (Wagon, Room,`

[1] A similar distinction of conceivability is also made in Philosophical Investigations [46], §193, in which Wittgenstein discusses the difference between a machine as a symbol and a machine in terms of its behaviour. Furthermore, there is evidence that Turing was also aware of Wittgenstein's position on some of these issues [39]. In short, the comment by Turing should not be seen as an attack on the method by which we arrive at the behaviour, rather it is an argument against the possibility of describing such behaviour at all.

> t) would result in the battery being removed from the room.

To put this in brief terms: We have a robot $R1$ and a task. $R1$ has a set of permissible actions. Each of these actions can be learned or formulated through logic, but all are meant to complete or work towards the completion of task. Dennett [10] discusses $R1$ does not understand all the important relations:

> Straight away it [R1] acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon.

Although the robot had a task and a set of actions, it had missed the relationship between the bomb and the wagon. As the creators understood, $R1$ had missed the relevancy of the context. So aside from the set of possible actions, each action should also be placed in a context such that: an action is desirable in context such that said action actually aligns and contributes to the completion of the task. Here of course the heart of the problem arises as the creators create another robot $R1D1$, that also deduces the relevant context for a specific action given a task.

> They placed R1D1 in much the same predicament that R1 had succumbed to ... It had just finished deducing that pulling the wagon out of the room would not change the colour of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon - when the bomb exploded.

Thus, we see the double-bind arise from the frame problem. We overshoot or undershoot in framing. In terms of relevancy, the machine overlooks certain factors, or it dies trying. What we can distil from this is the following:

The frame problem: How does a machine recognize the correct context and determines the relevant features in said context such that all actions result in or contribute to the accurate and correct completion of task?

This definition of the frame problem seems to revolve around relevancy. With this relation in our mind, we can see certain problems in AI in a different light. There are plenty of examples of AI systems making mistakes with moral implications, e.g. classifying the entrance to Auschwitz as being related to sport [21]. These problems may be addressed through some theory of value alignment, however, that does mean we need to sufficiently address this topic of relevancy.

The simplest way to understand the exact nature of this problem can be viewed through the perspective that Turing also proffered. The simple fact remains that such an AI system won't have been tested under literally all possible conditions, meaning there could and most likely will be mistakes. And those responsible may not be capable of overseeing what the consequences will be. If, however, we knew the relevant features and conditions, and could model those accurately, then knowing the trajectory of possible actions may be possible. It requires that we understand what context is involved, what the system highlights (puts emphasis on, gives importance to) and what the goal is specifically aligned to.[2] In this regard, such systems may still be brittle in novel circumstances, but if we apply them correctly and hem them in, then we should be able to use them effectively.[3]

The frame problem is an apt example when we desire to show the practical limits of value alignment theories without some interpreted notion of relevancy. Theories of value alignment may be helpful to decrease the scope of possible implementations, but it will practically still resort to some notion of relevancy to detail the context, teleology, and saliency. While there may be technical limits to the feasibility of certain aspects of relevancy (e.g. value trade-offs, or incompatible emphases, or simply intractable contextual scope), we do need to have a coherent and stable practice to assemble a reasonable grounding as to why certain choices were made throughout an AI lifecycle. Otherwise, we run the risk of maintaining a kind of anything goes attitude.

3.2 Relevant to whom?

By now, we should have a better understanding of the problem of value alignment. We need the right (moral) reasons for AI systems to function properly, but these are not given. Designers are likely not getting it right first time. This problem of course also counts for institutions and policies as well, yet we also know these can have potential benefits. Without taking away the work we do during implementation, we should start attaching more thought to the life cycle of a system once it is nestled in its context.

Yet there is one more major discrepancy in terms of relevancy which needs to be brought forth. These systems have to deal with the contention between different stakeholders, users, end-users, whom all may have different views which are incompatible perhaps on the level of context, perspective,

[2] Looked at from this perspective one can argue that the frame problem also essentially poses that value alignment in its ideal case is improbable if not infeasible. As we ourselves are also limited and may not have the capacity to really derive, formalize, or process what is relevant to a given situation. This is of course an abstraction of what value alignment is about—making better machines, which we can surely do by at least trying to incorporate these ideas and thoughts.

[3] This notion of correct application is also considered in the idea of a Moral Operational Design Domain (Cavalcante [5]).

or goals. These AI systems cannot easily entail to the monotheistic view of relevancy because it does not take into account the veritable jungle of opinions that stakeholders may have. What relevancy entails in value alignment is not merely a disambiguation of contexts and correctly specifying teleological aspects. Rather, it is about what a group of stakeholders think is relevant during the design rather than what is relevant. Not only could the stakeholders change over time, the intended use of such a system, or the context in which it sits, during design we also see the problem of relevance in terms of whom to invite to the table. Have we invited the relevant stakeholders?

One particular example that is interesting to note is recommendation systems for children. As recommender systems are mostly targeted at adults [28], yet children are also using these systems. As a stakeholder and likely their parents as well, designers need to consider more than merely the wants of the user. Instead, they may also need to incorporate very different dimensions, such as: educational, and developmental.

Yet, through contestation leading to adaptation, we could mend mistakes made during design. If individuals were capable of contesting an outcome and then having a kind of deliberation through or with the system, then some form of adaptation could be achieved, which could result in a better alignment of the system with the user.

4 Contestability and context

As mentioned in the introduction, we think contestability provides a good way of counteracting the problems of relevancy that we have thus far discussed. We also mentioned that—under the right circumstances—contestability could provide a better aligned system for users, while also giving practical insight to designers about how they ought to adapt their system while it is operating. To understand what it is we propose, we also must understand some part of contestability.

Literature on contestability is often focused on giving an inch of control to users when faced with automated decision-making [1]. For example, human intervention requires that one is able to contest the outcome of a decision before it is enforced. It can be seen as a kind of procedural justice, ensuring that participants have a voice in the matter. But human intervention may not be enough for some, it may also require that people can fully grasp the outcome, linking it directly to explainability [2]. If individuals are given an inkling as to why the outcome is what it is, they may be more substantive and understanding of what is going wrong during the decision-making [38]. In all, contestability focuses on the illegitimate or unjust decision that can arise from automated decision-making.

While this is certainly a good point to make, and in terms of user empowerment it is an interesting tool, there remains an open question as to what designers should do specifically after contestation has arisen. How should the system be changed or updated, and how should that be done? It is a question of operationalization. The main point of contestability, and how we would desire to present it, is that it provides a chance for realignment.

4.1 Contestation and framing relevant matters

The problem of relevancy given in Sect. 3.2 allows us to understand why contestability should come into play when dealing with relevancy. This does require that contestation actually leads to adaptation. It is our suggestion that such mechanisms for adaption need to be front and center after the implementation and that it is widely available, meaningful, and effective. This entails that we also look at alignment after first implementation. In this section, we propose an initial solution.

The main problem with proposing a solution is that it requires us to design and designate contextual features and goals, which may simply not align with the uncertain nature of the world. So, if the solution were to merely describe a theoretical framework that classifies what relevancy is, then it may cause the problem we wished to avoid in the first place. Designers may overshoot or undershoot in terms of our understanding of what is morally relevant in a given situation.[4] In fact, designers may limit themselves in terms of what is contextually available or what is acceptable for the telos of such machinery. At the other end of the spectrum, one may want to resort to requirements and guidelines, however those need to be followed effectively and truly, otherwise one runs into the problem of ethics washing. Furthermore, guidelines are often far too descriptive—rather than prescriptive—and can entail numerous things [20]. As Whittlestone et al. [45] also mention, certain AI perspectives are simply too broad and high-level to fill in the particulars. It seems both strict and lenient solutions cause another kind of version of the frame problem. In our opinion, neither a purely theoretical framework may suffice nor a mere set of guidelines. Rather, we need a practice—a taught method to become proficient with—which entices designers to think about their method of implementation in a specific way and guide the process to alignment itself.

[4] A nice example of these problems can be found in relevancy and communication [40], whereby the authors mention the difficulty of accessing the right amount of information given a situation. The problem with a purely theoretical framework of relevance is that it invites the same limitations that we wish to avoid in the first place.

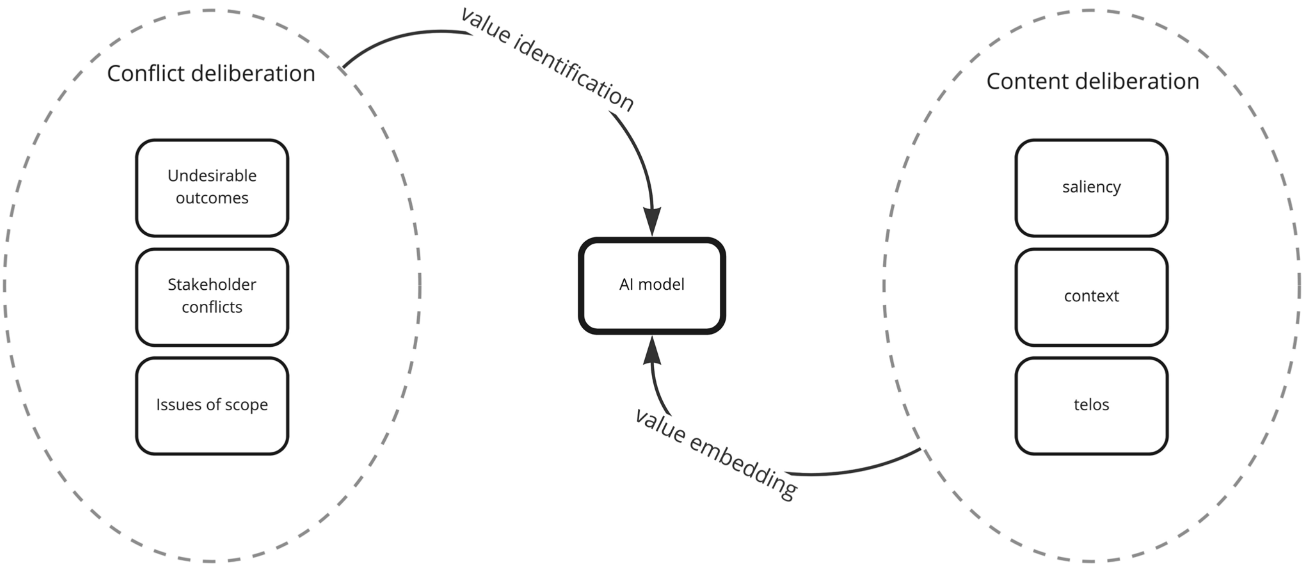

Fig. 1 Value identification and value embedding

What the frame problem already proffers to designers is the need for iteration.[5] The example given by Dennett entails that researchers need to go back to the drawing board and see that other problems pop up after implementation. Essentially every stakeholder that comes into contact with the system or is impacted by it needs to be able to contest the design of the AI such that it can morph into something more desirable [1].

The difficulty of pointing out flaws in AI systems is that these systems are overall effective. And contesting them may harm the effectivity of the model. Yet, these models are trained to work on general cases, meaning they often align well with the general cases one deals with. However, this also means that the edge-cases designers wish to avoid in may not arise at first inspection. In simple terms, reliability of a system is no necessary guarantee of safety [11]. There are bountiful examples to show that AI systems make unexpected turns and that designers lack the oversight both in its use and in the decisions such systems take [33].

The frame problem can show designers that any process of alignment is also a process of realignment. A simple example would be a navigation system which sends users through California wildfires [27]. There is a meaningful change to the context which we need to take into account, the problem is that it isn't easy to know when this happens. Nonetheless, what is open to us is conflict, or rather: contestation. A mistake by a machine happens through conflict, of what the machine puts out and what the users need, but that is merely an indicator after the fact. Value embedding and identification, we argue, should also be conflict-driven. The frame problem shows that framing matters as we cannot put our emphasis everywhere, yet that means it will most likely include a trade-off between certain ideals.

Thus, we believe that the constitutive elements of relevancy can feed into our ideas about what conflicts can arise during implementation and how to resolve them. These do not necessarily need to happen after a mistake has been made, but can also happen during discussions with stakeholders, the explication of implementation, argumentative structures for designing it in a particular way. In short, during each step of the life-cycle of an AI system, it could be that designers encounter a conflict. In Fig. 1 we give a quick overview of the two sides that can go into value-alignment strategies: content deliberation (e.g. how should we formulate the context? what should we optimize for?) and conflict deliberation (e.g. different stakeholders have differing opinions). These are two side of the same coin, as both embedding and identification of the relevant values rely heavily on formulation of the problem and the coinciding data collection to correctly formulate the root cause of the problem.[6]

[5] If we take a lesson from policy design instead, we can hearken back to Lindblom [24], he argues for the slow iterative process rather than leaps and bounds. The fact remains, Lindblom argues that in choosing policy (or in our case a specific type of implementation) is not made once and for all, it is successive because the objective and context is bound to change over time.

[6] The context of these problems are taken to be somewhat societal. These systems operate in some context that can influence or effect other agents (humans). Especially when these question become political e.g. when an algorithm which determines something about the height of the loans you can get, we inevitably come at the point where we must admit wickedness [7, 12].

Both during value identification and embedding, we can see that the constitutive elements of relevancy can be applied to think not merely in terms of accuracy, but rather in terms of outcomes and formulation. When we encounter obvious problems during this process of deliberation, we can start to understand where in our implementation certain measures must be taken. Depending on the problem at hand, a system designer can use one of these constitutive elements of relevancy to counteract or harmonize between conflicts. Only afterwards, when we have formulated the currently correct goal, with the correct data, and saliency, then accuracy comes into the picture as a measure of knowing whether the model is effectively trained.

These types of conflict and the tools to deliberate about them don't map specifically to any element of relevancy. They are rather possible locations where the problem resides. For example, when we encounter a problem, such as sexism in automated hiring processes [9] and racially biased data for risk assessment [32], it is unclear whether we optimized for the wrong thing (telos), whether our data was limited (context), or whether it simply had the wrong means to induce patterns (saliency). In fact, all three could be the case. For example, in the case of automated hiring, one may want to weigh the maintenance of current working culture (hiring based on similarity) versus a kind of openness and serendipity (hiring based on diversity), skewing this the wrong way shows that the translation from value to goal optimization went wrong.

Yet misappropriation of goals is not the only issue. Sometimes outcomes cannot be reached anyway. The most obvious example of this is fairness, as that heavily depends on the way the problem is scoped and formulated to even begin to understand the topic of fairness in a specific contextual setting. To say that the system must be "fair" is to pull off some equivocation, because fairness to a Rawlsian may mean something entirely different to your average communitarian. Most likely, these beliefs are incompatible to such an extent that optimizing for "fairness" is likely to result in unfair behaviour to some. To take the example of biased hiring, is sorting based on merit the way to achieve fairness? In The tyranny of merit [35], Sandel describes, merit may cause wildly unfair behaviour to arise. As it is bound to be influenced by socio-economic position and those who have the capacity to send their kids to all kinds of help, are going to end more or less on top. Yet, to leave it up to change may also be unfair. We only need to think in terms of probabilities and that small percentage that never gets hired once we implement such systems en mass.

After we have derived some conflict in whatever the stage of implementation, we need to know how the conflict is embedded in the system. This even may entail that we need to change the system completely, or to acknowledge that there is no solution for the problem that accommodate such reasons in a meaningful way. As we mentioned before, it is not necessarily clear where the problem may lie within the process of embedding, it is up to the designer at the point to potentially adjust and adapt one of the three elements that we described throughout this paper. See Fig. 2 for a quick overview.

The figure above describes the relation between conflict and adaption. The notification can be explained through stakeholders and the AI model itself. For example, concept drift is a way of showing that a user has drifted from its original interests, therefore the model needs to be updated. However, we also propose the fact that stakeholders can do this actively, as the AI models, notification of such conflicts may not be full proof (and open again to the frame problem). After notification of conflict, especially given the specific context, we do not want to look at this in an automated fashion because this may insert the original problem of the frame problem. Rather, the designer needs to play an active role in determining how to adapt the system.

Firstly, designers can adapt the context. This should be obvious, when an application is applying the wrong data, dataset, or set of propositions this can easily lead to skewed outcomes. Biased data is an often discussed topic in value alignment [30, 31]. Secondly, designers can adapt the telos, or what they optimize for. This stands in obvious connection to context. Yet, it starts with understanding that our outcomes are less descriptive and more moral than first meets the eye. Designers can view this multiple ways. Designers may gear an application towards something but leave out meaningful dimensions (like the bomb on the cart in the example of the frame problem), or they may misinterpret the situation such that these dimensions are seen as unimportant. Misfeaturization and emphasis, point us to the fact that such systems should not merely be built upon what the important facts are—but rather on what is needed and desirable here by the community at hand. Thirdly, designers can adjust saliency. For ML-approaches, this is the most difficult and time-consuming to disentangle. While in terms of relevancy this concept is about the noticeable, in AI practices designers need to understand this as induction or deduction of patterns. A particular kind of model can infer a particular kind of function, if that function disallows certain users to goods, or causes other harm, then changing the range of possibly deducible patterns may solve the problem at hand. It is easy to think about in terms of overfitting and bias. When designers introduce more bias, any inference pattern is less likely to adapt to an outlier, meaning that the pattern induced is more interested in the general whole. Yet, the necessity of fitting should be clear, we do not all act in the same manner, or need the same outcome from a system.

For symbolic approaches, we can see that the different kinds of logics that one can use and which types of deductions are acceptable is a far quicker choice to manipulate.

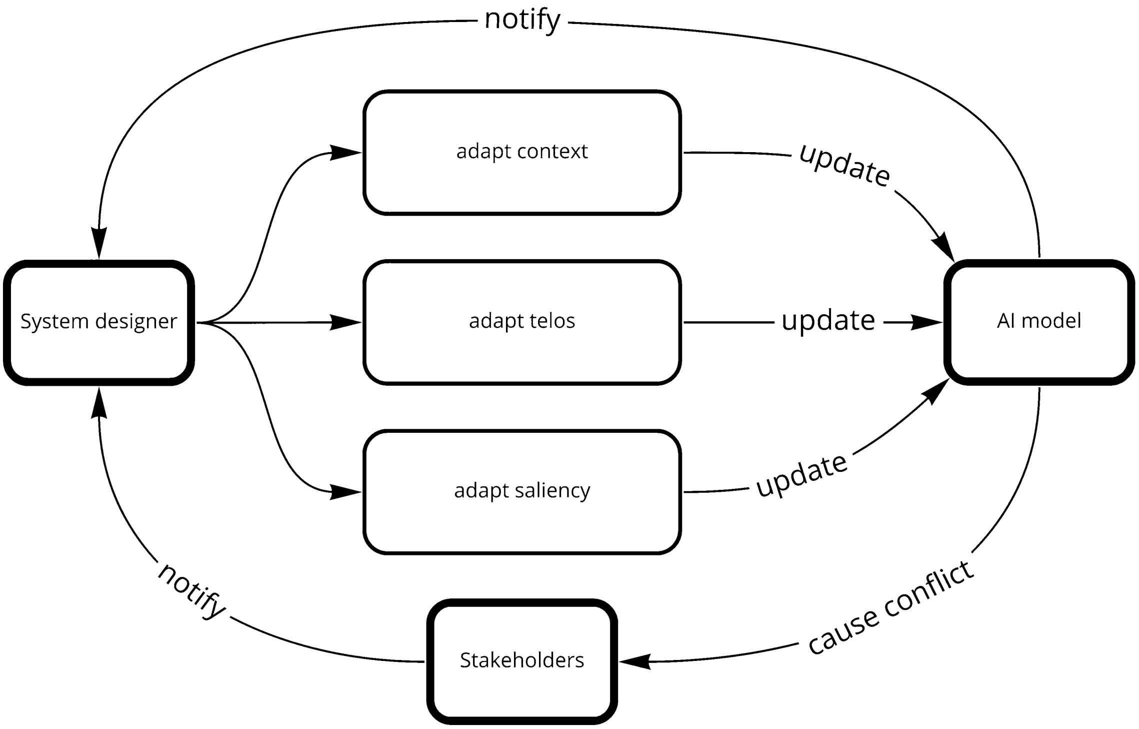

Fig. 2 Playing the updating game

One could also think about what kind of inconsistencies or contradictions that result from a dataset are acceptable and how they should be resolved. The entire approach of conflict acknowledgement and adaptation through the updating of context—telos—saliency is mostly similar within symbolic approaches, except for the fact that context is often painstakingly built up from countless propositions and swapping that out is likely to be such a time-dependent and consuming project that it equates to redoing the tool.

Relevancy through contestation invites designers to think about the means by which certain outcomes are achieved and the way in which a problem can become ingrained in the system when it happens to be misaligned. This way of thinking also shows designers a way how they can perhaps avert the problem. Just like coding etiquette, designers need to be taught in specific ways to make sure that even in larger teams with multiple designers, or working with legacy materials, can overcome a problem of misalignment in the future. So even when the system passes hands, it is still clear why the optimization strategy is what it is, why certain contextual features are scrapped or added, and why the pattern is inferred in one way and not another. Such documentation on the possible value conflicts may allow future designer(s) to re-align an application to the current day.

5 Conclusion

As we have seen, the openness that comes along with implementing value alignment theory can lead to misalignment during operationalization because what is relevant to a situation is far from obvious. The frame problem shows how difficult it is to get relevancy right, as it is far too easy to overshoot or undershoot in terms of deciding relevant factors. In all, we distinguish relevancy in value alignment mostly by what is thought relevant by a certain group of stakeholders, rather than say what actually is relevant. However, without means to adapt to new situations, this given relevancy is limited in a variety of ways. The context may be too limited, or the predesignation of the goal may be wrong. To effectively solve this, we argue that such systems should be built with an ingrained method to change the constitutive elements which are concerned with relevancy. We suggest that this could happen through contestability. This means that adaptability of systems needs to be in the system, together with feedback mechanisms that allow for meaningful contestation of individuals such that they can play an active part in the use of such a system. This creates a kind of update function that may prove worthwhile in approximating a desirable outcome for man and machine.

Funding This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alfrink, K., Keller, I., Kortuem, G., Doorn, N.: Contestable AI by design: towards a framework. Minds and Machines, Springer, Heidelberg (2022)
2. Almada, M.: Human intervention in automated decision-making: toward the construction of contestable systems. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, pp. 2–11 (2019)
3. Amodei, D., Chris, O., Jacob, S., Paul, C., John, S., Dan, M.: Concrete problems in AI safety. arXiv Pre-print arXiv:1606.06565 (2016)
4. Anderson, A.R., Belnap Jr, N.D., Michael Dunn., J.: Entailment, vol. II: the logic of relevance and necessity, vol. 5009. Princeton University Press, New Jersey (2017)
5. Siebert, C., Luciano, M.L., Lupetti, E.A., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., et al.: Meaningful human control: actionable properties for AI system development. AI Ethics **3**(1), 241–255 (2023)
6. Chow, S.J.: What's the problem with the frame problem? Rev. Philos. Psychol. **4**(2), 309–331 (2013)
7. Coyne, R.: Wicked problems revisited. Des. Stud. **26**(1), 5–17 (2005)
8. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., et al.: AI now report 2018, pp. 1–62. AI Now Institute at New York University, New York (2018)
9. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics, pp. 296–299. Auerbach Publications, Florida (2018)
10. Dennett, D.C.: Cognitive wheels: the frame problem of AI. In: Minds, Machines and Evolution: Philosophical Studies, pp. 129–150 (1984)
11. Dobbe, R.: System safety and artificial intelligence. In 2022 ACM conference on fairness, accountability, and transparency, 1584–4 (2022)
12. Dobbe, R., Gilbert, T.K., Mintz, Y.: Hard choices in artificial intelligence. arXiv Preprint arXiv:2106.11022 **300**, 103555 (2021)

13 Dunn, J.M., Restall, G.: Relevance logic. In: Handbook of philosophical logic, pp. 1–128. Springer, Heidelberg (2002)

14. Fisac, J.F., Gates, M.A., Hamrick, J.B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Dhruv Malik, S., Sastry, S., Griffiths, T.L., Dragan, A.D.: Pragmatic-pedagogic value alignment. In: Robotics research: the 18th international symposium Isrr, pp. 49–57. Springer, Heidelberg (2020)
15. Fodor, J.A.: Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In: Garfield, J.L. (ed.) Modularity in Knowledge Representation and Natural-Language Understanding, pp. 25–36. The MIT Press (1987)
16. Friedman, B., Hendry, D.G.: Value sensitive design: shaping technology with moral imagination. MIT Press, Massachusetts (2019)
17. Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A.: Value sensitive design and information systems. In: Early engagement and new technologies: opening up the laboratory, pp. 55–95. Springer, Heidelberg (2013)
18. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**(3), 411–437 (2020)
19. Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., & Dragan, A.: Inverse reward design. Advances in neural information processing systems, vol. 30 (2017)
20. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020)
21. Hern, A.: Flickr faces complaints over 'Offensive' auto-tagging for photos. The Guardian **20**, 2015 (2015)
22. Kim, T.W., Hooker, J., Donaldson, T.: Taking principles seriously: a hybrid approach to value alignment in artificial intelligence. J. Artif. Intell. Res.Artif. Intell. Res. **70**, 871–890 (2021)
23. Kim, T.W., Mejia, S.: From artificial intelligence to artificial wisdom: what socrates teaches us. Computer **52**(10), 70–74 (2019)

24 Lindblom, C.: The science of 'Muddling through.' In: Classic readings in urban planning, pp. 31–40. Routledge, London (2018)

25 McCarthy, J., Hayes, P.J.: Some philosophical problems from the standpoint of artificial intelligence. In: Readings in artificial intelligence, pp. 431–450. Elsevier, Amsterdam (1981)

26. Mizzaro, S.: How many relevances in information retrieval? Interact. Comput.Comput. **10**(3), 303–320 (1998)
27. Olsson, C.: Incident number 22. Edited by Sean McGregor. AI incident database. https://incidentdatabase.ai/cite/22 (2017)
28. Pera, M.S., Fails, J.A., Gelsomini, M., Garzotto, F.: Building Community: report on kidrec workshop on children and recommender systems at recsys 2017. In ACM Sigir Forum, ACM, New York (2018)
29. Peschl, M., Zgonnikov, A., Oliehoek, F.A., Siebert, L.C.: MORAL: aligning AI with human norms through multi-objective reinforced active learning. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 1038–1046 (2022)
30. Pratyusha, B.: World view. Nature **583**, 169 (2020)
31. Ribeiro, M.T., Sameer, S., Carlos, G.: "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, pp. 1135–1144 (2016)
32. Richardson, R., Schultz, J.M., Crawford, K.: Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. NYUL Rev. Online **94**, 15 (2019)
33. Russell, S.: Human compatible: artificial intelligence and the problem of control. Penguin, Westminster (2019)
34. Russell, S.: Artificial intelligence and the problem of control. Perspect. Digit. Humanism (2022). https://doi.org/10.1007/978-3-030-86144-5
35. Sandel, M.J.: The tyranny of merit: what's become of the common good? Penguin, Westminster (2020)
36. de Sio, S., Filippo, and Jeroen Van den Hoven.: Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI **5**, 15 (2018)
37. Saracevic, T.: The notion of relevance in information science: everybody knows what relevance is. But, what is it really? Synth. Lect. Inf. Concepts Retr. Serv. **8**(3), i–109 (2016)

38 Sarra, C.: Put dialectics into the machine: protection against automatic-decision-making through a deeper understanding of contestability by design. Glob. Jurist **20**(3), 20200003 (2020)

39. Shanker, S.G.: Wittgenstein versus turing on the nature of church's thesis. Notre Dame J. Form. Log. **28**(4), 615–649 (1987)

40. Sperber, D., Wilson, D.: Relevance: communication and cognition, vol. 142. Citeseer, New Jersey (1986)

41. Turing, A.M.: Computer machinery and intelligence. Mind **59**, 433–460 (1950)

42. Umbrello, S., Angelo, De, F.B.: A Value-Sensitive Design Approach to Intelligent Agents. In: Artificial Intelligence Safety and Security, pp. 395–409. Chapman; Hall/CRC (2018)

43. de PoelIbo., V.: Why new technologies should be conceived as social experiments. Ethics Policy Environ. **16**(3), 352–355 (2013)

44. de PoelIbo., V.: Design for value change. Ethics Inf. Technol. **23**(1), 27–31 (2021)

45. Whittlestone, J., Rune, N., Anna, A., Stephen, C.: The role and limits of principles in AI ethics: towards a focus on tensions. In proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, 195–200 (2019)

46. Wittgenstein, L.: Philosophical investigations. Wiley (2010)

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